Spring 1994

The complete nucleotide sequence of the sea lamprey (Petromyzon marinus) mitochondrial genome: Implications for the evolution of animal mitochondrial genome structure and rates of evolution in vertebrates

Woo-Jai Lee
University of New Hampshire, Durham

Order Number 9506423

The complete nucleotide sequence of the sea lamprey (*Petromyzon marinus*) mitochondrial genome: Implications for the evolution of animal mitochondrial genome structure and rates of evolution in vertebrates

Lee, Woo-Jai, Ph.D.

University of New Hampshire, 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

THE COMPLETE NUCLEOTIDE SEQUENCE OF THE SEA LAMPREY *(PETROMYZON MARINUS)* MITOCHONDRIAL GENOME: IMPLICATIONS FOR THE EVOLUTION OF ANIMAL MITOCHONDRIAL GENOME STRUCTURE AND RATES OF EVOLUTION IN VERTEBRATES

BY

WOO-JAI LEE

B.S. Chungbuk National University, Korea, 1987
M.S. Chungbuk National University, Korea, 1989
M.A. The City College, The City University of New York, New York, 1991

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy

in

Zoology

May, 1994

This dissertation has been examined and approved by:

Dissertation Director, Dr. Thomas D. Kocher
Associate Professor of Zoology

Dr. Ann C. Bucklin
Associate Professor of Zoology,

Dr. John J. Collins
Assistant Professor of Biochemistry

Dr. W. Huntting Howell
Associate Professor of Zoology

Dr. Peter F. Sale
Professor of Biology
University of Windsor, Canada

4/25/94

Date

To the memory of my lovely sister

ACKNOWLEDGEMENTS

I extend my deep-hearted appreciation and gratitude to my dissertation advisor, Dr. Thomas D. Kocher for inviting me to UNH, and, thereafter, for teaching, supporting and the care with which he guides my graduate study. I also would like to thank Dr. Kocher for his countless hours spent for review of each step and the final copy. I am indebted to my other committee members, Dr. Ann Bucklin, Dr. W. Hunt Howell, Dr. John Collins, and Dr. Peter Sale for providing a great opportunity to study fish biology and genetics, and for their thoughtful comments, suggestions, and encouragement from the beginning of this research. I also would like to thank Dr. Stacia Sower for the lamprey samples and Dr. Will Gilbert for his help at the step of data analysis.

My sincere thanks goes to Dr. Hung Sun Koh for his continuous concern and encouragement since my undergraduate time. Without his guidance, it would not be possible for me to step into my new career, nor open my eyes toward a higher goal.

This work could not have been completed without the help of many people. Thank you goes to Janet Conroy for her help during lab work for the last three years and Nicole Perna for her valuable time spent to read my rough draft and helpful comments. My special thanks goes to Dr. Marian Litvaitis for encouragement and good advice during my time in UNH, Jeremy Glasner for his friendship, and Diane Caporale for her help at the final step and her friendship.

I especially want to express my deep appreciation and gratitude to my family for their endless support and patience. My parents have sacrificed themselves to support me. My daughter, Gina, gave me refreshed stamina for more pipettings during the hot summer nights. I sincerely thank all of you!!!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

THE COMPLETE NUCLEOTIDE SEQUENCE OF THE SEA LAMPREY (*PETROMYZON MARINUS*) MITOCHONDRIAL GENOME: IMPLICATIONS FOR THE EVOLUTION OF ANIMAL MITOCHONDRIAL GENOME STRUCTURE AND RATES OF EVOLUTION IN VERTEBRATES

By

Woo-Jai Lee
University of New Hampshire, May, 1994

The complete nucleotide sequence of the sea lamprey (*Petromyzon marinus*) mitochondrial genome has been determined using purified, intact mtDNA. The lamprey mitochondrial genome is 16,201 bp in length, and contains genes for 13 proteins, 2 rRNAs, 22 tRNAs and 2 major non-coding regions. The genome displays several minor rearrangements but is basically colinear with other vertebrate genomes.

The sequence suggests that the vertebrate mitochondrial genomic organization was established at an early stage of vertebrate evolution. Comparisons with teleost and mammalian mitochondrial control regions demonstrate that some conserved sequence blocks have arisen recently. The overall base composition of the genome is similar to those of other chordate mitochondrial genomes. However the base composition at the wobble positions of four-fold degenerate codon families is strongly biased toward thymine.

Using amino acid sequences of nine other vertebrate and sea urchin

mitochondrial protein genes, the phylogeny and estimated rates of evolution among vertebrate lineages were obtained. The estimated rate of sequence divergence in the warm-blooded vertebrates is generally faster than in the cold-blooded vertebrates. However, this result may be attributed to the uncertainty of fossil records and rapid saturation for amino acid substitutions. Therefore, conclusions about relative rates of amino acid substitutions should wait for methods which account for the pattern of sequence saturation, as well as additional sequence data which may reveal the complete record of substitution in these genomes.

CHAPTER I

GENERAL INTRODUCTION ON EVOLUTION OF MITOCHONDRIAL DNA

Mitochondria, the energy-generating organelles in the cell, have their own genetic material and use their own genetic code to express it. Except for some hydroid species (Warrior et al., 1985), all metazoa have closed circular mitochondrial DNA (mtDNA) and the range of genome length is 13 kb to 40 kb. The genomic content is highly conserved throughout animal taxa. It contains genes for 12-13 proteins, 2 ribosomal RNAs (rRNA), 22-23 transfer RNAs (tRNA), 1-2 major non-coding regions, and short intergenic spaces between some genes. Over the last two decades, many studies have focused on the evolutionary mode and tempo of the molecule and it has become a useful tool for molecular evolutionary and population genetics. In this chapter, I review the mitochondrial genome structure and its evolution, the properties and applications of mtDNA, and finally the questions to be answered in the following work. Hereafter, mitochondrial DNA or genome refers to the animal mtDNA or genome, unless otherwise specified.

• **Abbreviations used:** mt, mitochondria(l); COI, COII, COIII, cytochrome oxidase subunits I, II, III; bp, base-pair(s); L (H)-strand, light (heavy)-strand; O_1 and O_2, first and second strand replication origin; D-loop, displacement loop; ND1 to 6, NADH dehydrogenase subunits 1 to 6; ATP6 and 8, ATPase subunits 6 and 8; Cyt b, cytochrome b; CSBs, conserved sequence blocks.

MITOCHONDRIAL GENOME CONTENT AND EVOLUTION OF ITS STRUCTURE

Genome Content

Animal mitochondrial genomes analyzed so far have almost identical genome contents. The mitochondrial genome carries genes for 13 proteins, 12S and 16S rRNAs, 22 tRNAs and 1-2 major non-coding regions. If present, most mitochondrial intergenic spacers are less than 10 bp. The total length of most genomes is 14-17 kb. The exceptions from the standard genome content are that two nematodes and the blue mussel have 12 protein-coding genes, and the latter has 23 tRNA genes (Hoffmann et al., 1992; Okimoto et al., 1992). The absence of introns, very short or no intergenic spacers and gene overlap result in the compact size of genome. One or two relatively long noncoding DNA segments carry most regulatory sequences. However, the sea scallop mtDNA sequence has expanded to about 40 kb, which is the longest metazoan genome identified so far, although the reason why it has such a long sequence and the content of the genome are unknown (Snyder et al., 1987).

The mitochondrial peptide-coding genes use a genetic code slightly different from the universal genetic code used in nuclear protein-coding genes. Two of the arginine (AGA, AGG), one isoleucine (AUA), and one stop (UGA) codons in the universal genetic code are used as termination, methionine, and tryptophan codons, respectively, in vertebrate mtDNAs. Moreover, some invertebrate mtDNAs use a more

modified genetic code than the vertebrate mitochondrial genetic code (Jacobs et al., 1988; Yokobori et al., 1993). The mitochondrially-encoded enzymes are a subset of the enzymes required for electron transport and oxidative phosphorylation. Other necessary enzymes are imported from the cytoplasm (Chomyn and Attardi, 1987).

All animal mtDNAs sequenced so far have 22 tRNA genes. One exception has been observed in the blue mussel mtDNA, which carries one more gene for tRNA-Met. The sequences of the mitochondrial tRNA genes are even more conserved than those of protein-coding genes. The conservation is probably necessary to maintain the ability to be folded into functionally important secondary structures. Post-transcriptional RNA editing of the tRNA-Asp anticodon has been observed in marsupial mtDNA (Janke et al., 1993). Otherwise, the anticodon sequences are identical in vertebrate mtDNAs. Most mutations in the stem regions of the secondary structure are compensatory base changes, and a few insertions or deletions (indels) are found in the loop sequences. Many nonstandard base pairs (A : C, and G : U) are observed in the stem regions. These unusual base pairings may cause the lower stability of the structure, and they are thought to be intermediate states to the completely compensatory changes (Brown, 1985).

The secondary structures of the two rRNA genes are highly conserved across taxa (Gadaleta, 1989). The patterns of base mutations are generally parallel to those observed in tRNA genes. Because of the longer sequence, larger inserts/deletions (indels) are found in the loop regions but some portions of rRNA genes are extremely

conserved.

One of the two major non-coding regions, the replication origin of the light strand, is located between genes of tRNA-Asn and tRNA-Cys in most vertebrate mtDNA molecules. It is about 30-40 bp in length and can be folded into a stem and loop structure, in which some compensatory base changes are observed between mammalian genomes. However, birds (Desjardins et al., 1990) and invertebrates (Jacobs et al., 1988; Hoffmann et al., 1992) do not have this segment. The other major noncoding segment is the control region containing displacement (D)-loop, which shows great length variation and sequence divergence. D-loop length makes the largest contribution to size variation among vertebrate mtDNAs (Moritz et al., 1987). Mammalian control regions display several conserved sequence blocks as well as highly variable sequence blocks. Invertebrate control regions contain a higher fraction of A+T, but have not yet been characterized in terms of structural evolution. The control region is the only segment containing tandem repetitive sequences in animal mtDNA, suggesting that some parts of the control region are free of selective constraints for structural conservation and therefore are more frequently subject to different genetic dynamics generating a complicated mode of evolution (Hoelzel, 1993).

Evolution of mitochondrial gene order

Despite the size variation and rapid rate of base substitution, changes of

genomic arrangement of mitochondrial DNA appear to be infrequent. The maintenance of gene order is probably attributable to the lack of major genetic exchanges by DNA turnover such as DNA slippage, unequal crossing over, and gene conversion (Moritz et al., 1987; Avise, 1991). The conservation of mitochondrial gene order may be informative about the origin and relationships of animal phyla (Brown, 1985). However, as complete mtDNA sequences accumulate, additional unique gene orders are being uncovered. It may be too early to determine the pattern of evolution in mitochondrial gene order. Among invertebrates, the complete or partial sequential orders of mtDNA genes of seven species representing four invertebrate phyla have been identified (Clary et al., 1985; Garesse, 1988; Smith et al., 1989; Jacobs et al., 1988; Hoffmann et al., 1992; Okimoto et al., 1992; Crozier et al., 1993).

Two nematodes (*Caenorhabditis elegans* and *Ascaris suum*) show an identical gene order and transcriptional polarity except for the transposition of an A+T rich region, but show quite different genomic organization from the other mtDNAs. The mtDNA of a mollusk (blue mussel) also represents a unique gene order, yet no other complete mtDNA sequence from this phylum is available to make a direct comparison. Restriction mapping of the sea scallop mtDNA, however, indicates that the genome is 40 kb in length, implying that the two mollusks underwent quite different mechanistic processes during evolutionary divergence. Comparison of mtDNAs from nematodes and the blue mussel shows that the transcriptional polarities are identical despite the altered gene orders, but neither shows similarity to the insect mtDNA in

gene order and transcriptional polarity.

In insects, two complete and one partial gene orders have been determined, of which two fruit flies show a couple of tRNA transpositions relative each other. In comparison with that of *Drosophila yakuba*, the gene order of honey bee (*Apis mellifera*) mtDNA also shows altered locations of 11 tRNA genes but the 8 genes of the 11 tRNA genes keep the same transcriptional polarity. The rest of the genes are in the same positions with an identical transcriptional polarity.

Among echinoderms, only sea urchin mtDNA sequence is completely identified and it shows no affinity with any other group of metazoan mitochondrial genome organization. However the genomic organization within the phylum is very informative to clarify the relationships of echinoderm classes (Smith et al., 1993). The patterns of gene order group sea urchin and sea cucumber together, while brittle stars and sea stars form a separate group. Among vertebrates, the mitochondrial gene order is a highly conserved feature except for a few minor changes (Anderson et al., 1981; Roe et al., 1985; Árnason et al.,1991; Chang et al., 1994). Transpositions of tRNA genes and a duplication followed by deletions have been observed in the marsupial and chicken mtDNAs respectively (Janke et al., in press; Desjardins et al., 1990). In chordate mtDNAs sequenced so far, the transcriptional directions are basically identical.

The mitochondrial genomic order plainly shows a basic arrangement within each phylum, in spite of many rearrangements at the intraphylum level during

evolutionary divergence. Rearrangements can be explained by simple inversions or duplications followed by deletions. Nevertheless, more data are necessary to see whether or not the gene order reflects the evolutionary history of each phylum.

CHARACTERISTICS OF MITOCHONDRIAL DNA

Rate of Base Substitution

From comparisons of mtDNA sequences of closely related taxa, the rate of base substitution in mitochondrial DNA has been estimated. It is evident that the average divergence rate of mtDNA is 5-10 times faster than that of single-copy nuclear DNA (Brown et al., 1979). This increased rate has been attributed mainly to deficient or absent repair mechanisms, high rate of DNA damage, and relaxation of functional constraint (Cann and Wilson, 1983; Wilson et al., 1985).

The evolutionary rate of mtDNA sequences among mammals has been intensively studied (Wilson et al., 1987; Hasegawa et al., 1993). The average rate of mtDNA divergence in primates is approximately 2% per million years per lineage (Wilson et al., 1985) whereas a 5-6 times slower rate of evolution has been reported in sharks (Martin et al., 1992). At the amino acid level, the cold-blooded animals have a 6 times slower rate of evolution than do the warm-blooded vertebrates (Adachi et al., 1993). The fast evolutionary rate in warm-blooded animals has been attributed to an increased rate of damage caused by the higher metabolic rate (Martin et al., 1993).

The rate of base changes is not homogeneous across mitochondrial genes. The tRNA and rRNA genes evolve more slowly, while some sites in protein-coding genes

accumulate much variation in a short period of time, because of the different selective constraints on the different types of genes. The tRNA genes form thermally stable secondary structures, in which the stem regions of tRNA genes appear to accumulate base changes linearly and slowly (Kumazawa et al., 1993). In contrast, in the small ribosomal gene, the long single-stranded regions evolve slowest (Vawter et al., 1993), suggesting that the large ribosomal gene may have the same base substitution pattern along the whole molecule. Slowly evolving portions of RNA genes might be more useful for resolving deep-branches of animal phylogeny, while the more rapidly evolving portions of the 16S gene have been useful at the population and species level of divergence. The third codon positions of protein-coding genes are generally thought to be nearly free from selective constraint allowing the accumulation of much variation quickly. The first and second codon positions evolve more slowly. The control region (non-coding region) shows heterogeneity in the rate of base changes. The portions of the control region corresponding to regulatory sequences evolve extremely slowly, and the rest of the control region which carries no apparent function, evolves more rapidly (Brown, 1985; Saccone et al., 1991). In the mitochondrial genome, base substitutions predominate over indels except in some segments of the control region and ribosomal RNA genes, where frequent long indels and tandem repeats are observed (Johansen et al., 1990; Wilkinson et al., 1991). Protein-coding genes show very few indels, possibly due to the fact that most indels can disrupt the protein conformation and function. The rapid rate of base changes and

few indels facilitate the use of protein-coding sequences for studies between closely related taxa, perhaps relationships between or within species, while the slowly evolving genes like tRNA genes are more suitable for analysis of distantly related animal taxa. More direct comparisons in early diverged vertebrates and invertebrates are needed to establish the universality of evolutionary rate found in higher vertebrates.

Through comparisons between mtDNA sequences from closely related species, it is well known that, in animal mtDNAs, transitions (A $\leftrightarrow$ G, C $\leftrightarrow$ T) greatly outnumber transversions (A $\leftrightarrow$ C and T, G $\leftrightarrow$ C and T). For example, in rat cytochrome oxidase subunit II (COII) gene, the ratio is 8.0 : 1 (Brown et al, 1982). Moreover, the rate of transition, for instance A $\rightarrow$ G or G $\rightarrow$ A, is not always equal resulting in unequal base composition (Tamura, 1992; Perna and Kocher, 1994). When comparing mtDNA sequences from taxa above the species level, we detect less divergence than expected, because more than one base substitution at the same position (multiple hits) have occurred, hiding previous base changes. Thus, in protein-coding genes, the third positions of codons are saturated quickly. The multiple hits need to be accounted for, by an appropriate method, to get a reliable phylogenetic tree (Kimura, 1980).

The pattern of base substitution within a gene is often heterogeneous. Despite the usual concordance between base composition and codon usage, the base changes at the third positions of four-fold degenerate codons do not always conform with the

base composition of the mitochondrial genomes. Nonrandomness in the rate of base changes and codon usage is perhaps due to different mutational spectra between or within mtDNAs (Brown, 1985; Perna and Kocher, 1994).

Maternal Inheritance

Since Hutchinson et al.'s (1974) report, based on the results of backcrossing hybrids to the male parent, it has been known that animal mtDNAs display a strong maternal inheritance, in contrast to biparental inheritance of nuclear DNA. After a decade, several investigations confirmed the previous results (Lansman et al., 1983; Gyllensten et al., 1985). The average number of paternal mitochondria in the next generation was estimated to be no more than 10^{-3}. More recently, paternally inherited mtDNA molecules in progeny were detected using the polymerase chain reaction (PCR) (Gyllensten et al., 1991). The frequency of paternal inheritance was still less than 10^{-4}.

Because of the quantitative difference in the mtDNA molecules contributed by each parent [the number of mtDNA molecules in a mature oocyte and in a sperm cell are about 10^5-10^8 and 50-100 copies respectively (Avise, 1991; Michaels, 1992)], the ratio of maternally inherited mtDNA among the progeny is not inconceivable even under an assumption of biparental inheritance. Therefore, the quantitative majority of maternal mtDNAs in the next generation does not prove a maternal inheritance of mtDNA.

Since the genotypes of heteroplasmic mtDNA in hybrid zones, where parental species with highly divergent genotypes live together, showed highly similar sequences, it has been argued that mitochondrial heteroplasmy might arise from mutations in maternal lineages, rather than from paternal inheritance (Avise, 1991). Moreover, 95% of heteroplasmy found in white sturgeon was not attributed to the paternal leakage of mtDNA, but to DNA slippages associated with replicational processes, although the paternal inheritance was suspected as a source for the rest of the observed heteroplasmy (Brown et al., 1992). Thus, the mode of maternal inheritance and the effects of paternal leakage in mtDNAs still remains questionable.

Paternal leakage, if it occurs, has important consequences for interpretations of mtDNA data. It has been thought that the mtDNA heteroplasmy observed in various animal taxa (Mulligan et al., 1989; Mignotte et al., 1990; Kondo et al., 1990; Wilkinson et al., 1991) originated from the small number of paternally inherited mtDNA molecules. Heteroplasmy of mtDNA caused by paternal leakage may lead to a difficulty when genealogical analysis is carried out with an assumption of maternal inheritance (Wilson et al., 1985; Gyllensten et al., 1991).

Nevertheless, the dominance of the maternal genotype in the next generation is one of the properties that makes mtDNA a sensitive marker for detecting population structure and female-mediated gene flow (Birky et al., 1989 and 1991).

Lack of Genetic Exchanges

Unlike nuclear DNA, animal mtDNA is small and compact. It does not have introns, long spacers between genes, or transposable elements. Genetic exchange in nuclear DNA has been well demonstrated. The typical mechanisms leading to genetic exchange are recombination, unequal-crossing over, transposition by mobile elements, and duplication. By these mechanisms, genes can move around consequently altering the evolutionary rate of the genes significantly. Because of the altered evolutionary rate, phylogenetic information is not as sensitive as unchanged evolutionary rate (Jacobs et al, 1988). The conserved structures of mtDNA across animal taxa are mostly attributed to the lack of mechanisms for genetic exchanges (Moritz et al., 1987). Thus, the lack of the genetic exchange enhances the use of mtDNA as a tool for inferring the evolutionary history of animals.

The continuous discovery of tandem repeated sequences in the control region (D-loop) of vertebrate mtDNA suggests the presence of more complicated mechanistic processes than previously thought (Mignotte et al., 1990; Avise, 1991). DNA turnover, at least in the control region of the animal mtDNA, might generate and maintain the repeats. The existence of genetic recombination in animal mtDNA has not been demonstrated, although it is suggested that DNA turnover might occur in the control (D-loop) region (Olivo et al., 1983; Hoelzel, 1993). However it seems to be evident that there is no major genetic exchange in the rest of the animal mitochondrial genome, which maintains a highly conserved genomic organization and relatively

constant rate of evolution.

In addition to the rarity of genetic exchanges within mtDNA (intragenomic exchanges), it is unknown whether or not genetic exchange takes place between mtDNA and nuclear DNA (intergenomic exchanges), although mtDNA uses proteins and RNAs originating from nuclear DNA (Clayton, 1987). A number of previous studies found mtDNA-like sequences in nuclear DNA (Fukuda et al., 1985; Zullo et al., 1991; Smith et al., 1992). However, no DNA segment originating from nuclear DNA has been found in mtDNA. The apparent lack of transfer of genetic material from the nuclear genome further contributes to the stable rate of evolution in mtDNA.

APPLICATIONS OF MITOCHONDRIAL DNA SEQUENCES

Population Genetics

The characteristics of mtDNA described above have facilitated the extensive use of mtDNA for population and evolutionary biology. Especially, the rapid evolutionary rate and maternal inheritance of mtDNA make mtDNA a sensitive marker of genetic differentiation between subpopulations, female-mediated gene flow, and other population events (Birky et al., 1983; Takahata et al., 1985)

In contrast to biparental inheritance of nuclear DNA, the nearly completely uniparental inheritance of mtDNA ensures mtDNA homoplasmy in the cell. If there is no sexual bias in a population, the effective population size for mtDNA will be nearly equal to the number of females (N_f), four-fold smaller than the effective population size for nuclear DNA (N_f versus $2N_e$). Consequently, because of a lower effective migration rate ($m_e = \alpha m_f + \beta m_m$, for α and β, see below), meaning lower possibility of gene flow between populations, mtDNA will more sensitively show the subdivisions of a population than nuclear DNA (Takahata et al., 1985; Birky et al., 1989 and 1991). The formulas developed by Birky et al., (1989) clearly show that the gene diversity of a subpopulation (G_s) of mtDNA reaches equilibrium more rapidly than that of nuclear DNA: $t_{1/2}=\ln 2/(2m_e + 1/N_{e\infty})$, where t indicates time to reach equilibrium, m and m_e are total and effective migration rate, and $N_{e\infty}$ shows effective

number of mitochondrial genes. $N_{\infty} = N_m N_f /(\alpha^2 N_m + \beta^2 N_f)$, where α and β indicates the fraction of maternal and paternal inheritance respectively. If mtDNA is inherited nearly maternally, β will be very small, making N_{∞} large and consequently the time (t) shorter. In fact, many studies using mtDNA have showed high level of polymorphism within species, indicating subpopulation structures from human to copepods (Johnson et al., 1983; Bermingham et al., 1986; Ball et al., 1988; Danzmann et al., 1991; Nolan et al., 1991; Bucklin et al., 1992). MtDNA sequences, for instance, revealed more subdivisions in a population which were undetected by allozyme study, a nuclear measure of genetic differentiation (Carr et al., 1991). On the other hand, if a population has passed through a bottleneck in size or has a high level of dispersal of females, the mtDNA will lose diversity, and show homogeneity in the population (Wilson et al., 1985; Avise et al., 1986; Moritz et al., 1987).

It is important to keep in mind that since the assumption of a mode of strict maternal inheritance of mtDNA may not be true, the detectable amount of paternal inheritance of mtDNA can obscure the interpretation of data obtained under the assumption of complete maternal inheritance.

Phylogenetic Studies

The rapid rate of evolution and straightforward record of mutations in maternal lineages have facilitated the use of mtDNA as a tool for phylogenetic studies, especially for closely related taxa. Before the emergence of DNA sequencing

technology in the 1980's, most phylogenetic studies were based on comparisons of mtDNA restriction fragment polymorphism. The large-scale length polymorphism found among closely related taxa (Bermingham et al., 1986; Billington et al., 1991), however, made the method less reliable. Because of their greater sensitivity, DNA sequences have been used more often for phylogenetic inference since the late 1980's. Obtaining DNA sequences from various sources, such as ancient and museum samples, has become easier and faster due to the refinement of sequencing techniques such as PCR and automated sequencing (Kocher, 1992). Moreover, the sequences may provide closer insight into understanding the rudimentary mode of mtDNA evolution. For instance, the pattern of base composition and codon usage can be uncovered only by using DNA sequences.

Differences in genomic organization can be used for phylogenetic information between phyla, sometimes between classes (Smith et al., 1993). A limited number of genomic organizations, however, has been determined. The locations of tRNAs vary more often, but it is unknown whether or not these differences provide phylogenetic information for closely-related taxa. The homogeneity in evolutionary rate and small-scale indels within most of mitochondrial protein-coding genes can provide useful phylogenetic information for relatively closely related taxa (Moritz et al., 1987; Normark et al., 1991). As an example, nucleotide sequence from the cytochrome b gene, one of the most frequently sequenced mitochondrial genes, provides reliable information for phylogeny (Irwin et al, 1991). Mitochondrial tRNAs, which show a

more conserved mode of sequence divergence, have also been used for deeper branches of animal phylogeny (Kumazawa et al., 1993). For ribosomal DNA sequences, since the selective constraints vary at the different sites along the secondary structure, character weighting might be necessary (Wheeler et al., 1988; Gatesy et al., 1992). As a result of the universality of primers designed from a certain part of mitochondrial 16S gene, the 16S sequences have been widely used for population studies as well as phylogenetic studies for both closely and distantly related taxa (Xiong et al., 1990; Peer et al., 1990; Bucklin et al., 1992; Gatesy et al., 1992). The control region, the fastest evolving segment in the mtDNA molecule, has a more complicated mode of evolution, possibly due to DNA turnover (Hoelzel, 1993). Thus using the whole control region for phylogenetic analysis might not be appropriate, and, moreover, choosing a segment within the control region for phylogenetic analysis above the species level should be done with particular care (Lee and Kocher, 1994).

Sequence data can be used to infer phylogenetic history using one of the three major tree reconstruction methods; parsimony, distance, or maximum likelihood. The maximum parsimony method, which is one of the most frequently used methods for phylogenetic studies, is to search for a tree that requires the minimum number of base substitutions. In other words, the most parsimonious tree has the shortest tree length. The same character at the same site across taxa under study carries no phylogenetic information in the maximum parsimony method. Therefore, only informative sites that have at least two kind of nucleotides in DNA data represented in at least two taxa

each are taken into consideration. Because there often exist several trees of equally short length, it is often impossible to choose just one most parsimonious tree. The parsimony method also assumes that all lineages are homogeneous in evolutionary rate. It is observed that parsimony is very sensitive to the violation of this assumption (Hasegawa, 1993). Although parsimony is not based on other substantive assumptions about evolutionary processes, it considers all base substitutions equal. However, the pattern of mtDNA base substitutions is more than simply base changes. Consequently, if other evolutionary processes such as transition / transversion bias and multiple base changes at the same sites are under consideration, it is difficult to use the parsimony method (Jin and Nei 1990). Nevertheless, if the data have a large number of informative sites, parsimony is thought to be a robust tree reconstruction method.

Distance methods begin by calculating all pairwise differences between taxa. DNA sequence differences are transformed into a distance matrix using any one of a number of different algorithms which correct for multiple hits. According to the algorithm used, the distance matrix may vary, affecting the phylogenetic topology. Next, one of the several tree constructing methods uses the distance matrix to generate a tree. For example, UPGMA, which was originally developed for numerical taxonomy, uses the distance between two taxa, assuming they have the same evolutionary rate. Lake's distance and neighbor-joining distance methods, however, are not based on this assumption, and use different algorithms to reconstruct phylogenetic tree with the matrix (Lake, 1987; Saitou et al., 1987).

The maximum likelihood method calculates the probability of the transformation from one nucleotide to another across all sequences, including the sites with the same bases, and yields likelihood values for a particular tree (Felsenstein, 1981). The tree which maximizes the likelihood value is chosen. The probability of base transformation may be different according to the evolutionary model employed. Despite the fact that this method theoretically has less room for error in estimating relationships between lineages with unequal or unknown evolutionary rates (Felsenstein, 1981 and 1988), it is still unfeasible to handle large nucleotide data sets with maximum likelihood method due to the complicated and computationally expensive algorithms. Recently, amino acid sequences have been used instead of nucleotide sequences for the maximum likelihood method (Hasegawa, 1993).

IMPORTANT QUESTIONS TO BE ANSWERED IN THIS DISSERTATION AND IN FUTURE WORK

First, the evolutionary origin and historical timing of the standard vertebrate mitochondrial gene order, and the structural evolution of the mitochondrial control region are to be explored. Among invertebrates, the patterns of gene order are informative for distantly related taxa (above class level). The mtDNA gene order in vertebrates is colinear from bony-fish to human except for a couple of simple rearrangements. However, the pattern of evolution in the control region appears to be more complicated and less phylogenetically informative, despite a number of intensive studies on the mammalian control region (Hoelzel et al., 1991; Mignotte et al., 1990). Getting clearer insight to the evolutionary timing of the common genomic arrangement and the pattern of evolution in the control region are important to understanding the evolution of vertebrate mtDNA. Toward this end, the complete mtDNA sequence of sea lamprey and the structure and evolution of several fish control regions will be presented in chapters III and IV, respectively.

Second, the phylogeny of distantly related vertebrates is to be reconstructed using amino acid sequences of all mitochondrial-encoded protein genes. The maximum likelihood method is supposed to be the most reliable tree construction method regardless of the rate of evolution among lineages (Hasegawa, 1993). Through comparisons of trees made by different methods, it may be possible to support a

method for tree reconstruction and provide closer insight into understanding evolutionary rates among vertebrate lineages. Also the pattern of mitochondrial amino acid substitutions will be uncovered. These will be discussed in chapter V.

In addition to the objectives disscussed in this dissertation, as a long term goal, the evolutionary relationships of early vertebrates and the origin of tetrapods are to be inferred. The phylogeny of distantly related vertebrates (perhaps, above class level) has been well established mostly by the means of fossil records and comparative analysis of morphological characters. For some cases, molecular data have also been used (Field et al., 1988; Meyer et al., 1990). However, many questions about the phylogenetic relationships within classes still remain. One of the remaining problems is the branching order of early diverged vertebrates. There are two orders of living jawless fish (agnatha), and the other more derived jawed vertebrates (gnathostomes). The two groups can be distinguished by the presence or absence of jaws. Other characters, however, make it more difficult to clarify the relationships between agnathans as well as between agnatha and gnathostomes mainly due to the difficulty of determining which characters are ancestral or derived (Hubbs et al., 1971). A number of recent reports are also in conflict (Jamieson., 1991; Stock et al., 1992; Forey et al., 1993). Despite data from a number of molecular studies (Gorr et al., 1991; Hillis et al., 1991; Meyer et al., 1992) the origin of tetrapods is still in debate (Gorr et al., 1993) as a result of unknown relationships among fishes (jawless fish, shark, teleost, lungfish, and coelacanth) and the affinity of one of the fishes to

tetrapods. All previous studies used three fish species (teleost, lungfish, and coelacanth) and relatively short DNA or nuclearly encoded protein sequences. Adding more taxa like lamprey and shark and using mtDNA sequences, in which the rate of evolution and the function of genes have been well determined, will provide higher confidence in the reconstructed tree.

CHAPTER II

ISOLATION, CLONING, AND SEQUENCING OF MITOCHONDRIAL DNA

For studies using mitochondrial DNA, a number of methods such as hybridizations, restriction fragment length polymorphisms (RFLP), and applications of DNA sequences, are available. For most of those methods, isolation of pure mtDNA, which is very time consuming and labor intensive, is unnecessary. For example, the development of the polymerase chain reaction (PCR) has made sequencing an interesting gene or segment of DNA feasible without requiring isolation of intact mtDNA. However, the purification of mtDNA is still necessary if a study requires a probe covering a large part of the mitochondrial genome, or when primers in PCR synthesize multiple products because of the impurity of template DNA. Sequencing a large part of a mitochondrial genome conventionally requires pure mtDNA for the purpose of cloning. In this chapter, I will describe the methodological basis of mtDNA isolation, cloning, and sequencing used for sequencing the complete sea lamprey mitochondrial genome.

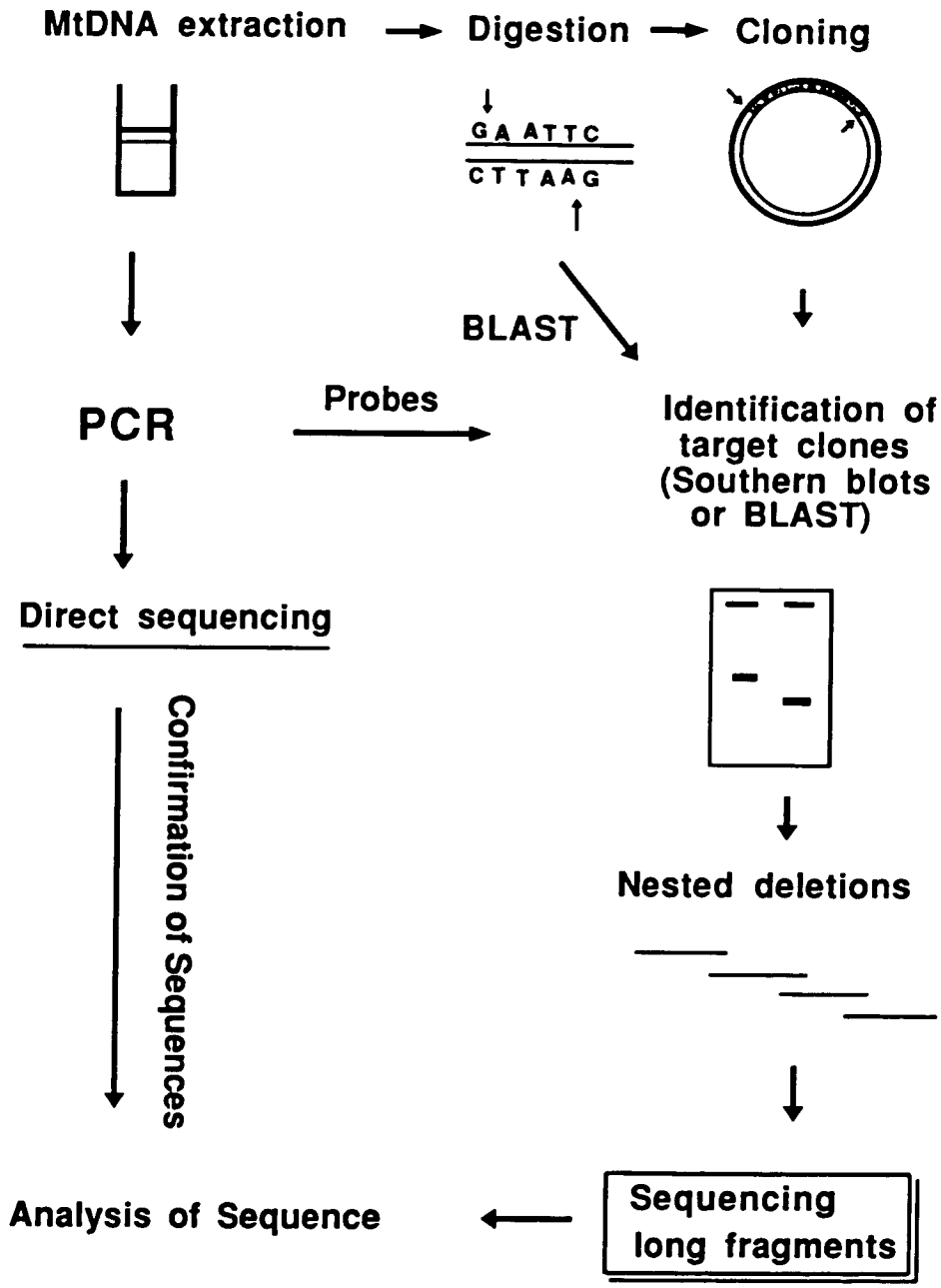
SAMPLE SELECTION AND STORAGE

The selection of tissues for mtDNA isolation from live organisms depends on the ratio of mitochondrial DNA to nuclear DNA. More active cells usually contain more mitochondria, the energy generating facilities in the cell, to provide enough energy for smooth cell metabolism. For instance, one mature oocyte carries 2.6×10^5 copies of mitochondria, while some somatic cells contain 100-fold fewer copies (Michaels et al., 1982). Based on this fact, tissues from liver, gonad, heart, or kidney have been widely used for mitochondrial DNA studies (Lansman et al., 1981; Dowling et al., 1990). Although tissues from oocytes or gonads of fishes are frequently used (Chapman et al., 1984; Wilson et al., 1990), the liver tissue from sea lamprey yielded more supercoiled mtDNA than mature eggs.

In a number of previous reports (Lansman et al., 1981; Chapman and Powers, 1984; Wilson and Tringali; 1990), the use of fresh tissue was highly recommended. Because of damage caused by freezing and thawing procedures on the membranes of cell organelles, the yield of intact mitochondrial DNA from frozen tissue is significantly reduced. However, in cases where freezing is unavoidable under certain circumstances, quick freezing using liquid nitrogen or dry ice is a way of reducing the degree of damage (Dessauer et al., 1990). Since the collection of sea lamprey in fresh water is feasible only during the spawning season (April-June in the East-coast

region), use of fresh tissue, other than during the period of reproduction, is very restricted. Therefore, I stored the tissues in a deep freezer (-80°C) after quick freezing using liquid nitrogen. For other samples such as winter flounder and yellowtail flounder, fresh liver tissues were used. I could not obtain intact mtDNA from uncarefully frozen tissues. Moreover, the quantity of intact mtDNA recovered from the quick frozen tissues was usually less than 50% of that from fresh tissues. Overall schematic diagram of the procedures used for mtDNA sequencing is shown in Figure 2-1.

Figure 2-1. Overall schematic diagram of procedures used for sequencing of the lamprey mitochondrial genome. Intact mtDNA was extracted by ultracentrifugation in CsCl-Etbr gradient. The restriction enzyme used was *EcoRI* and the cloning vector was *pBluescript II*. The primers for PCR were universal or newly designed. The conditions of PCR were 93°C, 50°C, and 72°C for denaturation, annealing, and extension respectively and 30-35 thermal cycles. For Southern blot, non-radioactive labelling method was used. For BLAST, sequences from ends of the 4 fragments were sent to the Genbank.



MITOCHONDRIAL DNA EXTRACTIONS

Homogenization of Tissues

All the following steps were carried out at 4°C unless otherwise specified. About 10-15 g of fresh or partially thawed frozen liver or egg tissues were minced with razor blades prior to homogenization. Each 5 g of tissue was ground in 15 ml of the homogenization buffer described in Lansman et al. (1981) (0.21 M Mannitol, 0.07 M Sucrose, 0.05 M Tris-HCl, pH 7.5, and 3 mM $CaCl_2$), with a motor driven tissue homogenizer [Wheaton, or Tissumizer (Tekmar)]. Using Tissumizer was much more convenient as well as more productive. After 2-3 strokes of grinding, 0.5 M EDTA was added, making a final concentration of 20 mM, and the homogenate was mixed well by inverting the tubes. The tubes were kept on ice until low speed centrifugation.

Low Speed Centrifugations

The purpose of this step is to get rid of cell debris and nuclei. Although most previous studies spun samples at 700-800 x g, two consecutive centrifugations at 2300 rpm (550-600 x g) in a Hermle removed nearly all debris. Furthermore when 700-800 x g was applied, the yield of supercoiled mtDNA was minimal, perhaps because of the precipitation of mitochondria at that speed. The supernatant was carefully transferred

to clean tubes after each centrifugation. The final supernatant from 2-3 low speed centrifugations was much clearer and more aqueous.

High Speed Centrifugation and Lysis of Mitochondria

The final supernatant from previous low speed centrifugations was transferred to 15 ml tubes and centrifuged at 20,000 x g for 30 min to pellet mitochondria. The precipitated mitochondria were resuspended in 1 ml of STE buffer (O.1 M NaCl, 0.05 M Tris-HCl, pH 8.0, and 0.01 M EDTA, pH 8.0) and lysed by adding 100 µl of 20% SDS. After incubation at room temperature for 10 min or at 37°C for 5 min, 200 µl of CsCl-saturated water was added to the sample and mixed well. The mixture was placed on ice for 30 min and incubated at 4°C overnight. Heavy precipitation of debris was visible the next morning. It was obvious that the removal of a greater portion of the SDS and remaining debris was necessary in order to see distinctive nucleotide bands after ultra-centrifugation. The mixture was centrifuged at 2,000 x g for 10 min, which eliminated most of the SDS and any remaining cell debris (Dowling et al, 1990). Usually the supernatant was viscous after precipitation. When it was too viscous, the solution was diluted with TE buffer (0.01 M Tris-HCl and 0.5 mM EDTA, pH 8.0).

Ultracentrifugation in CsCl-ethidium Bromide Gradient

Ultracentrifugation through a CsCl-ethidium bromide gradient has been widely

used to isolate supercoiled DNA from linear DNA, based on the density of DNA molecules intercalated with ethidium bromide (Lansman et al., 1981; Sambrook et al., 1989). Pure mitochondrial DNA molecules from the liver tissues of sea lamprey, winter flounder, and yellowtail flounder were obtained using ultracentrifugation as follows:

80 µl of ethidium bromide (10 mg in 1 ml of TE buffer) and 0.90 g of solid CsCl were added to every 1 ml of suspension. The density of the solution after adding CsCl was adjusted to the range of 1.45-1.50 g/ml, but lower than 1.55 g/ml: The higher density placed the DNA bands near the top of the tubes that had a filling capacity of 3.9 ml, which made it difficult to rescue intact mtDNA. The adjusted DNA solution was loaded into ultracentrifuge tubes with a 22G syringe and centrifuged at 90,000-100,000 rpm at 20°C for 16-24 hours using a Bekman TL-100 table top ultracentrifuge.

Two DNA bands were visible under UV illumination (Figure 2-2). The lower band, containing mtDNA, was usually more obscure and narrower, and located about 5 mm below the nuclear DNA band. The band was not visible when the amount of mtDNA was too small. The mtDNA band was drained with a 22G syringe through the top of the tubes.

To remove ethidium bromide, the mtDNA solution was purified with water-saturated butanol four times until the pink color was gone. CsCl was eliminated with a fast dialysis method (Centricon-30, Amicon) at the recommended speed. Four or

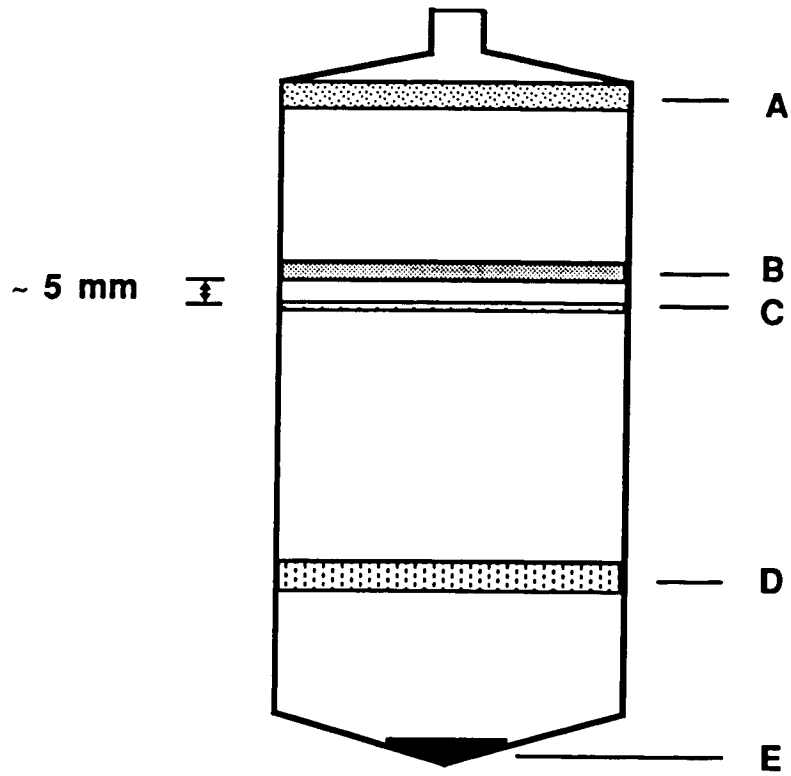


Figure 2-2. Diagram of visualized ultracentrifuge tube containing CsCl - ethidium bromide gradient with mitochondria lysed with 1% SDS. The centrifuge carried out at 100,000 rpm for 18 hrs separated the cell components.
A, extra Et-Br with SDS; B, nuclear DNA and linearlized mtDNA; C, Supercoiled intact mtDNA; D, glycogen (more seen, when eggs were used); E, cell debris and RNA pellet.
The distance between two nucleotide bands was about 5 mm or less.

five centrifugations, adding 1 ml of sdH_2O after each spin, gave me pure enough mtDNA for further analysis. Since small traces of CsCl inhibit subsequent enzymatic reactions, the complete elimination of the salt was very important. The DNA concentration was determined with a DNA fluorometer and aliquoted mtDNA was stored at -20°C for a couple of months until I was ready for the next step. The yield of mtDNA from 15 g of fresh liver was about 5 µg.

Other Methods

Besides the ultracentrifugation in CsCl-ethidium bromide gradients, a number of alternative protocols for the isolation of mtDNA have been developed (Powell and Zúñiga, 1983; Chapman and Powers, 1984; Welter et al., 1989; Wiesner et al., 1991). Of those, I used the Chapman and Powers method with some modifications. Basically, I followed the standard protocol described above until the point of the high speed centrifugation, which was replaced by a sucrose-gradient centrifugation described in Chapman and Power, 1984.

Fifteen percent sucrose-TEK solution (50 mM Tris, 10 mM EDTA, 1.5 KCl, pH7.5) was very carefully layered under the supernatant from the low speed spins, using a syringe, and centrifuged at 18,000 x g for 30 min. The distinctive boundary remained between the two solutions, the high density sucrose phase and low density sucrose homogenate. The supernatant was transferred to clean tubes and mitochondria were pelleted by spinning at 20,000 x g for 30 min.

After lysis of the pelleted mitochondria as described in the previous section, instead of 20% SDS, using 10% non-idet (NP-40, Sigma) making a final concentration of 1%, cell debris and intact nuclei were pelleted by centrifugation at 12,000 x g for 30 min.

The supernatant was thoroughly mixed with an equal volume of TE saturated phenol to denature proteins. I shook the mixture for longer than 5 min and the suspension was centrifuged at 6,000 rpm for 10 min to remove denatured proteins. This phenol extraction was repeated twice and extracted with an equal volume of phenol/chloroform until the suspension was clear. Finally, using chloroform, I removed the remaining phenol from the solution. The aqueous phase was mixed with twice the volume of cold absolute ethanol (-20°C) and incubated at -20°C for two hours. The final centrifugation at 12,000 x g for 20 min precipitated the nucleic acids. The pellet was dried in a vacuum dryer and resuspended in 100 µl of TE buffer containing DNAase free RNAase A (1µl of 5,000 units from Promega). The DNA was kept at -20°C.

DIGESTION OF SUPERCOILED MTDNA AND CLONING

Digestion with Restriction Enzymes

The purified mtDNA was digested with four restriction enzymes (*EcoRI*, *HindIII*, *BamHI*, and *PstI*). In order to get fragments with appropriate sizes for cloning, only enzymes recognizing 6 nucleotides were chosen. For each digestive reaction, about 0.5 µg of purified mtDNA, 1 µl of 10x digestion buffer supplied by the manufacturers, and 1-2 units of the restriction enzyme in 10 µl total volume were incubated in a 37°C water bath overnight or for a minimum of 4 hours. The completion of digestion was checked on 1.0% agarose gels. 2.5 µl of ethidium bromide (10 mg/ml) was added to every 50 ml of melted 1% agarose to see the DNA fragments under UV light. Agarose gels in 1x TAE buffer (0.04 M Tris-acetate, 0.001 M EDTA, and 1.14 ml glacial acetic acid) were run for 2-3 hours at 57 volts. Of those enzymes, *EcoRI* generated 4 fragments with suitable sizes for cloning into plasmids (*pBluescript II SK*) while *HindIII, BamHI, and PstI* generated 8, 1, and 5 fragments with very small or very large sizes. The four *EcoRI* fragments were approximately 6 kb, 4.5 kb, 3.5 kb, and 3 kb respectively (Fig. 2-3).

Cloning

Preparation of Cloning Vectors The cloning vector (*pBluescript II SK*) in *E.coli*

Figure 2-3. MtDNA fragments digested with EcoRI and stained with EtBr. Lanes 1 and 3 represent the size standards of λ phage DNA/Hind III and 1 kb fragments respectively. The lengths of the EcoRI fragments in lane 2 are 6, 4.5, 3.5, and 3 kb respectively.

(*XL-1 Blue*) host was donated by Dr. Cathy Tugmon.

A maxiprep method (Sambrook et al., 1989) with some modification was used for large-scale plasmid preparation. In a 15 ml sterile culture tube, 10 ml of autoclaved LB broth (20 g/l), ampicillin making a final concentration of 50 µg/ml, and 1-2 µl of the vector stock, were mixed and incubated in a shaking incubator at 250 rpm for 16 hours. The total culture was poured into 500 ml of autoclaved LB broth containing the same concentration of ampicillin as in the previous overnight culture, and tetracycline with a final concentration of 12.5 µg/l. This second amplification was carried out overnight under the same conditions and the supercoiled pBluescript was isolated using an alkaline lysis method as follows.

The bacterial cells were collected from the second overnight culture by centrifugation at 1,800 x g for 15 min at 4°C. The bacterial pellet was resuspended in 100 ml of 4°C STE buffer (0.1 M NaCl, 10mM Tris-Cl pH 8.0, and 1 mM EDTA pH 8.0). The cell suspension was precipitated again using the same conditions as above. 18 ml of GTE solution (50 mM glucose, 25 mM Tris-Cl, pH 8.0, and 10 mM EDTA, pH 8.0) were added to the bacterial pellet. 2 ml of freshly made lysozyme (10 mg/ml in 10 mM Tris.Cl, pH 8.0) were added and mixed by inverting the tube. 40 ml of freshly prepared 0.2 N NaOH/1% SDS solution were added into the tube and gently mixed. The lysis of bacterial cells was completed by incubation at room temperature for 10 min. 20 ml of cold neutralization stock solution (60 ml of 5 M potassium acetate, 11.5 ml glacial acetic acid, and 28.5 ml dH_2O) were added and

shaken, until no more distinct liquid phases were visible. After incubation at room temperature for 10 min, the bacterial lysate was centrifuged at 1,800 x g for 15 min at 4°C. Some floating materials were removed by filtering the supernatant through 4 layers of cheesecloth, and 0.6 volume of isopropanol was added to precipitate the nucleic acids. After incubation at room temperature for 15-20 min, the clots of nucleic acids were visible. The precipitation of DNA was finished by centrifugation at 3,000 x g for 15 min at room temperature. A big DNA pellet was rinsed with 70 % ethanol twice and air dried. The DNA pellet was dissolved in 3 ml of TE, pH 8.0. The supercoiled plasmid DNA was isolated by ultracentrifugation in a CsCl-ethidium bromide gradient as described in a previous section.

The intact plasmid DNA was stored at -20°C. 2-3 µg of plasmid DNA were digested with the same enzyme as used for mtDNA (*EcoRI*). The completeness of digestion was checked by loading a small aliquot in a 1% agarose gel. The digested plasmid DNA was purified by phenol/chloroform extraction and ethanol precipitation. After vacuum drying, the DNA pellet was resuspended in 10 µl of TE buffer.

Preparation of Competent Cells Preparation of a frozen stock of competent cells for cloning has been recommended (Dagert et al., 1979; Hannhan, 1985). The benefits of making a frozen stock are time savings and rapid accessibility. After treatments with $CaCl_2$, the bacterial cells are so fragile that the competent cells should be handled very carefully (Sambrook et al., 1989). I found that the efficiency of transformation

using frozen competent cells was not noticeably different from using fresh cells.

All cell cultures were performed at 37°C unless otherwise specified. In a sterile 15 ml culture tube containing 10 ml of LB broth only (without ampicillin), bacterial cells (*E.coli* XL-1 blue) were inoculated from the surface of a frozen stock. The cells were incubated overnight (for about 16 hours) with agitation. The culture was diluted with new LB broth (1:20) in a Pyrex 50 ml glass culture flask and incubated with shaking at 250 rpm, until the optical density (O.D.) reached 0.5 at 660 nm. It usually took 3-4 hours. However after the O.D. reached 0.3, the cells grew quickly. Therefore, the O.D. was checked every 20 min, after it reached 0.3. The culture was divided into 4 orange-capped 15 ml tubes and kept on ice for 20-30 min. After centrifugation at 1,200 x g for 10 min at 4°C, the supernatant was completely poured off. The bacterial pellets were resuspended in a half volume of ice cold 50 mM $CaCl_2$ and the clumps of cells were dissolved by gentle pipetting. The tubes were placed on ice for 20 min with occasional swirling. After another spin at 1,200 x g for 10 min at 4°C, the $CaCl_2$ was thoroughly drained off. The pellets were resuspended with 1/15 volume of cold 50 mM $CaCl_2$ made 15% in glycerol and placed on ice for an hour. 200 µl aliquots in sterile 1.5 ml microtubes were frozen quickly with liquid nitrogen and stored in -80°c.

Ligation, Transformation, and Plating The ligation reaction was carried out with both pBluescript and mtDNA digested by the same restriction enzyme. Instead of

isolating the mtDNA fragments from the digestion, the total reaction was used for ligation. However, the plasmid DNA was purified with phenol/chloroform after digestion. Ligation reactions were performed at 15°C overnight. Total volume of the ligation reaction was kept under 20 µl to maintain an optimal DNA concentration. Two times more mtDNA than plasmid DNA was used in order to maximize the ligation efficiency. About 2 µg and 1 µg of digested mtDNA and plasmid DNA respectively, 2 µl of 10x ligation buffer, and 3 units of ligase were mixed in a 20 µl reaction. After overnight incubation, the reaction was diluted 4-fold with TE buffer, and stored at -20°C.

200 µl of frozen competent cells were thawed on ice prior to transformation. 30-40 ng of recombinant DNA and 200 µl of competent cells were mixed in a prechilled 12 x 75 mm centrifuge tube. The mixture was placed on ice for 30 min before heat shock in a 42°C water bath for 90 seconds. The tube was placed on ice for 3 min to select transformants. 500 µl of LB broth without ampicillin was added to the transformants. The tube was incubated in a 37°C shaking incubator for 30 min, but no longer. Meanwhile, reagents for plating were prepared.

Petri dishes containing 10-15 ml of LB agar, and two antibiotics (ampicillin, 50 µg/ml; tetracycline, 12.5 µg/ml) were prepared before ligation was started. On the surfaces of LB agar plates, 100 µl of 100 mM IPTG and 40 µl of 2% X-gal, were spread with a glass rod. The rod was ethanol and flame sterilized before and after each use. Besides the transformants, I plated 100 µl of competent cells as a control.

After incubation of plates inverted at 37°C for 18 hours, blue and white colonies appeared and no colonies were observed in the control plate. The white colonies were selected for further analyses.

Identification of Recombinants Containing mtDNA

Prior to sequencing, the recombinants carrying the target DNA segments must be identified. The two basic methods for clone identification will be described in this section.

Recovery of Recombinant DNA

The recombinant DNAs from the white colonies were obtained using a short alkaline-lysis method (Sambrook et al., 1989). A commercially available kit (Magic miniprep, Promega) which made it simple to get more supercoiled DNA, was also used.

3 ml of LB broth containing the two antibiotics were inoculated with a single white colony, and incubated overnight. Two consecutive centrifugations of 1.5 ml tubes at the top speed of a table top microcentrifuge allowed me to collect bacteria from 3 ml overnight cultures. The bacterial pellet was resuspended in 200 µl of GTE buffer. 300 µl of freshly made 0.2 N NaOH/1% SDS were added to the resuspended bacterial cells and the contents were mixed by inverting the tubes. After incubation at room temperature for 5 min, 300 µl of cold neutralization solution were added, and

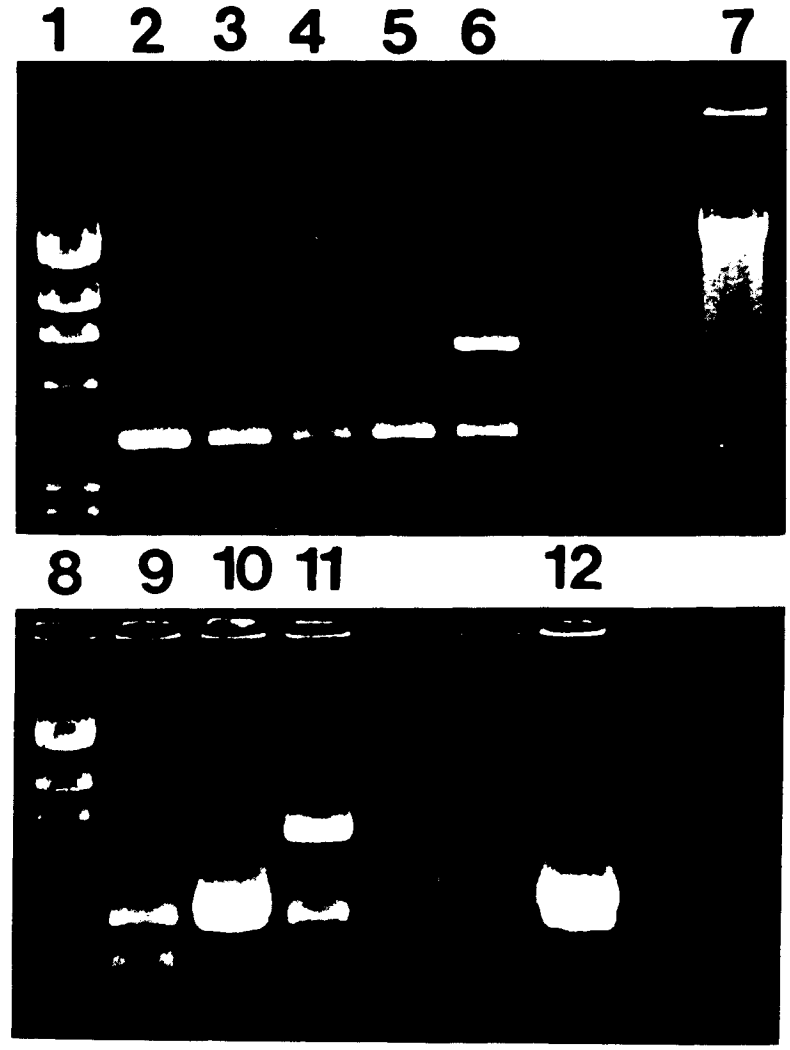
the mixture was placed on ice for 5 min. The cellular debris was removed by centrifugation for 10 min, and the supernatant was transferred to a clean tube. 1 µl of DNAase-free RNAase A eliminated most of the RNA from the supernatant. After adding an equal volume of cold isopropanol, the DNA was precipitated by centrifugation at the top speed for 15 min. The DNA pellet was dried and resuspended in 50 µl of TE buffer. To determine the sizes of inserted DNA, 1 µl of miniprepped DNA was run on a 1% agarose gel (Fig. 2-4). The colonies carrying inserts corresponding to the size of the target DNA were stored at -80°C for further analyses.

Identification of mtDNA Inserted in Plasmid

Two methods were employed to identify the recombinants carrying the mitochondrial fragments. The first method that I used was Southern blotting, using PCR products amplified from the lamprey mtDNA as probes. Fragments for which no probes were available were identified by homology to mitochondrial DNA sequences in the Genbank data base.

Southern Blot For use as hybridization probes, three conserved segments of the lamprey mtDNA were amplified using PCR. Parts of genes for Cytochrome b, large ribosomal RNA, and Cytochrome oxidase subunit I, were chosen because of the slow rate of base substitution among animals. I substituted Digoxigenin-11-dUTP

Figure 2-4. Gel picture of min prepped DNAs from recombinants. Lanes 1 and 8 are DNA size markers of λ DNA/*Hind III* fragments. Lane 2 contains plasmid DNA only. All recombinant DNAs were digested by *EcoRI*. Colonies in lane 3 through lane 5 contain no insert DNAs bigger than the plasmid DNA. The upper bands in lanes 6 and 11 represent insert DNA fragment of about 6 kb, and the upper bands in lanes 10 and 12 together show another 3.5 kb foreign fragment. Lane 7 contains undigested sea lamprey mtDNA.



(Boehringer Mannheim), non-radioactive labeling material for dTTP (Emanuel, 1991) in the standard PCR protocol (Kocher et al., 1989). The completion of labeling was easily determined, for the labelled PCR products migrated slowly on a 1% agarose gel (Fig. 2-5). Before transferring the recombinant DNA from a 1% agarose gel to a nylon membrane, the DNA was denatured in 0.5 N NaOH/1.5 M NaCl and neutralized in 1.0 M Tris-HCl, pH 8.0/1.5 M NaCl for 30 min respectively. The DNA was transferred to a nylon membrane by capillary method, in 10x SSC buffer (1.5 M NaCl, 150 mM sodium citrate, pH 7.0) overnight. The membrane was washed in 5x SSC buffer for 1 min at room temperature and placed on a piece of blotting paper (Whatman 2mm). The membrane was baked in an oven at 80°C for an hour to fix the DNA on the membrane. In a self-sealing plastic bag, the membrane proceeded to prehybridization in hybridization solution [5x SSC, 1.0% blocking reagent (Boehringer Mannheim), 0.1% N-lauroylsarcosine, and 2% sodium dodecyl sulfate] for 2 hours. The membrane was transferred to a clean plastic bag and the same amount of fresh hybridization solution and labelled DNA probe were added into the bag. Both prehybridization and hybridization were done at 65°C with gentle agitation. The probe was denatured by boiling for 10 min, before it was added to the bag. Hybridization was performed overnight and the membrane was rinsed twice with 2x wash solution (2x SSC/0.1%SDS) for 5 min per wash. The membrane was washed twice again with 0.5x wash solution at 65°C. Hybridizing fragments were visualized using colormetric detection using Genius system kit (Boehringer Mannheim). All steps followed the

Figure 2-5. Agarose gel (1%) picture of amplified probes labelled with non-radioacitive material, Digoxigenin-11-dUTP. Each volume of PCR was 100 µl and amplified with universal primers. 5 µl of each reaction was run. Lane 1 is for negative control of the PCR. From left, each two lanes were amplified DNAs from partial COI, 16S, and Cyt b genes. The second lanes in each pair represent the labelled PCR products, which migrate slightly slower than normal PCR products. The clones carrying inserts were hybridized with amplified probes.

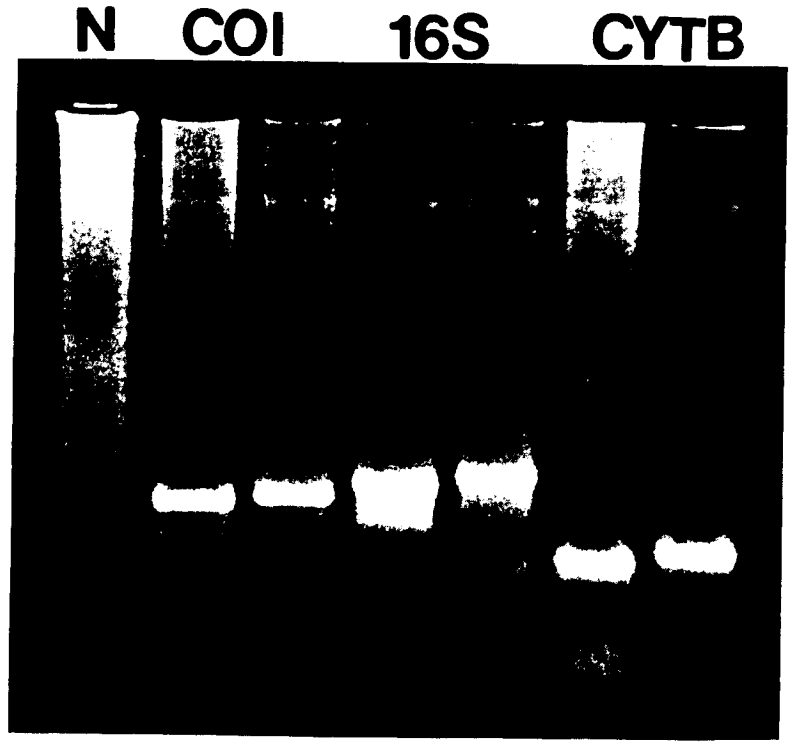
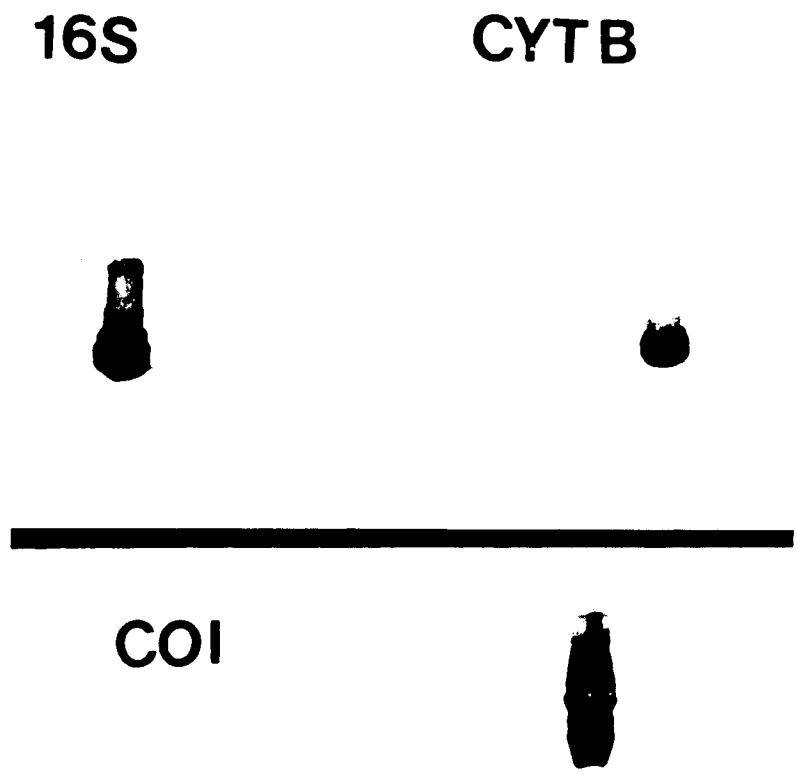


Figure 2-6. Southern blots using amplified non-radioactive (Digoxygenin-11-dUTP) probes to DNAs of clones carrying the same size of inserts as mtDNA fragments. Positive clones were hybridized with probes from 16S, Cyt b and COI genes in panels 1, 2 and 3 respectively.



manufacturer's manual. The results are presented in Figure 2-6. One fragment was positively detected with the COI probe and another was identified with two probes from 16S and Cyt b genes. Thus two of four fragments were identified by Southern blotting.

Basic Local Alignment Search Tool (BLAST) searches Because of the unavailability of probe for genes in the other two unidentified clones, recombinant clones with the same size insert as the target mtDNA fragments were selected after minipreps described in a previous section. Both ends of each clone were sequenced using M13 universal primer. About 250-400 nucleotides from each end of clones were obtained and the sequences were sent to Genbank through the ethernet network to find sequences with the best similarity (Fig. 2-7). The clones with a high similarity to other animal mtDNAs were chosen and sequenced completely. It is concluded that because more and more sequence data accumulates, the BLAST search (Altschul et al., 1990) as an identification method is very useful and sensitive, especially when limitations make it difficult to employ other methods.

Figure 2-7. An example of BLAST results. 310 bp of one insert DNA isolated from a recombinant was sent to the Genbank to search sequences with the most similarity. All sequences in the result are mitochondrial sequences, suggesting the query sequence is a mitochondrial sequence. Furthermore, from the result, it is evident that the source of the query sequence is not from a mammalian mtDNA which can be incorporated the most easily during mtDNA isolation procedures.

National Center for Biotechnology Information (NCBI)
Experimental GENINFO(R) BLAST Network Service (Cruncher)

• NUCLEOTIDE SEQUENCE DATABASES

Sequences producing High-scoring Segment Pairs:		High Score	Smallest Poisson Probability P(N)	N
gb|L29771|ONHMTCG	Oncorhynchus mykiss mitochondrion co...	454	1.3e-28	1
gb|X52392|MIGGX	Chicken mitochondrial genome	225	5.1e-20	2
gb|M91245|CRQMTGENOM	Crossostoma lacustre mitochondrion, ...	253	5.2e-19	2
gb|X17662|MTGMGEN5	Atlantic cod mitochondrial DNA for t...	332	6.0e-19	1
emb|X61010|MICCCG	C.carpio complete mitochondrial geno...	340	7.8e-19	1
gb|X61010|MICCCG	C.carpio complete mitochondrial geno...	340	7.8e-19	1
gb|M34496|CHKMTTGHA	Chicken mitochondrial His-tRNA gene.	179	3.7e-07	1
gb|V00659|MIHL45	Gibbon mitochondrial genome fragment...	196	1.1e-06	1
gb|Z29573|DVMTGNME	D.virginiana mitochondrial DNA compl...	177	4.9e-05	1
gb|M22656|TARMTTGH	Mitochondrion T.syrichta NADH-dehydr...	176	5.4e-05	1
gb|V00711|MITOMM	Mouse mitochondrial genome.>gb|J0142...	166	0.00041	1
gb|L07095|MUSMTHYPA	Mus domesticus hydrophobic protein m...	166	0.00041	1
gb|L07096|MUSMTHYPB	Mus domesticus hydrophobic protein m...	166	0.00041	1
gb|V00675|MIPY45	Orangutan (P. pygmaeus) mitochondria...	164	0.00055	1
gb|M22655|SAIMTTGH	Mitochondrion S.sciureus NADH-dehydr...	149	0.010	1
gb|X72004|MIHGCG	H.grypus mitochondrial sequence, com...	146	0.020	1
gb|X03297|HSMTLIK3	Human nuclear mitochondrial-DNA-like...	140	0.045	1
gb|J01394|BOVMT	Bovine mitochondrion, complete genome.	140	0.062	1
gb|V00654|MIBTXX	Complete bovine mitochondrial genome.	140	0.062	1
gb|M22650|MACMTTGHA	Mitochondrion M.mulatta NADH-dehydro...	137	0.10	1
gb|X63726|MIPVDNA	P.vitulina mitochondrial DNA, comple...	137	0.11	1
gb|M22653|MACMTTGHC	Mitochondrion M.fascicularis NADH-de...	134	0.17	1
gb|X02890|MIXLG	Xenopus laevis complete mitochondria...	117	0.996	1
gb|M10217|XELMTCG	X.laevis mitochondrial DNA, complete...	117	0.996	1

>gb|L29771|ONHMTCG Oncorhynchus mykiss mitochondrion complete genome.
Length = 16,660

Minus Strand HSPs:

Score = 454 (125.4 bits), Expect = 1.3e-28, P = 1.3e-28
Identities = 142/206 (68%), Positives = 142/206 (68%), Strand = Minus

```
Query:   213 ATAGTTTATACAAAACATTAGATTGTGAGTCTAATAAAGAAGGTTAAAATCCCTCTGCCT 154
            ||||||||  ||| ||||||||||||||| ||||| ||   ||||||||||||   |  |
Sbjct: 12727 ATAGTTTAACCAAGACATTAGATTGTGATTCTAAAAATAGAGGTTAAAATCCTCTTATCC 12786

Query:   153 GCCGAGAGGGGCAAGGCAGCACTAAGAACTGCTAATTCTTTTCCCCTGAGGTTCAACTCC 94
             |||||||      |     |  |   |||||||||  | | ||||    ||| || ||
Sbjct: 12787 ACCGAGAGAAATCTGTTGATAACAGAGACTGCTAATCTTCTGCCCCCTCAGTTAAATTCT 12846

Query:    93 ACAGCCCTCTCGAGCTTCTAAAGGATAAGCAGCAATCCGCTGGCCTTAGGTGCCACCAAT 34
               |  | |||| ||||||||||||||      |||| ||| ||||||  |||  | |
Sbjct: 12847 GTGGTTCACTCGTGCTTCTAAAGGATAATAGCTCATCCATTGGTCTTAGGAACCAAAACT 12906

Query:    33 CTTGGTGCAAATCCAAGTAGAAGCTA 8
            ||||||||||||||||||| |||||
Sbjct: 12907 CTTGGTGCAAATCCAAGTAGCAGCTA 12932
```

```
 Score = 162 (44.8 bits), Expect = 9.4e-12, Poisson P(2) = 9.4e-12
 Identities = 42/54 (77%), Positives = 42/54 (77%), Strand = Minus

Query:   302 CACCCGTGAACACCTACTTATACTTATACACATAGCCCCTATTATCCTTCTCAT 249
             |||||| ||||||||||||||||  || | ||| |   ||| ||| |||||||| ||
Sbjct: 12632 CACCCGAGAACACCTACTTATTATTCTGCACCTCATCCCAATTGTCCTTCTAAT 12685


>gb|X52392|MIGGX Chicken mitochondrial genome
           Length = 16,775

  Minus Strand HSPs:

 Score = 225 (62.2 bits), Expect = 4.4e-09, P = 4.4e-09
 Identities = 73/108 (67%), Positives = 73/108 (67%), Strand = Minus

Query:   221 AAGCACGCATAGTTTATACAAAACATTAGATTGTGAGTCTAATAAAGAAGGTTAAAATCC 162
             | |||  |||||||||  | ||||||||||||||||| ||||| ||     ||| ||  |
Sbjct: 12862 ATGCAAACATAGTTTAACCCAAACATTAGATTGTGATTCTAAAAATAGGAGTTTAACCCT 12921

Query:   161 CTCTGCCTGCCGAGAGGGGCAAGGCAGCACTAAGAACTGCTAATTCTT 114
             |  ||   |||||| || |        ||  ||||||||||||||| |
Sbjct: 12922 CCTTGTTCGCCGAGGGGAGGCCCAAGCCAGCAAGAACTGCTAATTCCT 12969

 Score = 211 (58.3 bits), Expect = 5.1e-20, Poisson P(2) = 5.1e-20
 Identities = 47/53 (88%), Positives = 47/53 (88%), Strand = Minus

Query:    63 AGCAATCCGCTGGCCTTAGGTGCCACCAATCTTGGTGCAAATCCAAGTAGAAG 11
             ||||||||| ||| ||||||  ||||| ||||||||||||||||||||| |||
Sbjct: 13016 AGCAATCCGTTGGTCTTAGGAACCACCTATCTTGGTGCAAATCCAAGTAAAAG 13068


>gb|M91245|CRQMTGENOM Crossostoma lacustre mitochondrion, complete genome.
           Length = 16,558

  Minus Strand HSPs:

 Score = 253 (69.9 bits), Expect = 1.9e-11, P = 1.9e-11
 Identities = 73/101 (72%), Positives = 73/101 (72%), Strand = Minus

Query:   135 GCACTAAGAACTGCTAATTCTTTTCCCCTGAGGTTCAACTCCACAGCCCTCTCGAGCTTC 76
             ||| |||| ||||||||| ||| |  ||   |||| ||||||   || |||||| |||||
Sbjct: 12719 GCAATAAGCACTGCTAATACTTATAACCCACGGTTAAACTCCGTGGCTCTCTCGTGCTTC 12778

Query:    75 TAAAGGATAAGCAGCAATCCGCTGGCCTTAGGTGCCACCAA 35
             ||||||||||      ||||  ||| ||||||  |||  ||
Sbjct: 12779 TAAAGGATAACAGCTCATCCATTGGTCTTAGGAACCAAAAA 12819

 Score = 205 (56.6 bits), Expect = 5.2e-19, Poisson P(2) = 5.2e-19
 Identities = 53/68 (77%), Positives = 53/68 (77%), Strand = Minus

Query:   209 TTTATACAAAACATTAGATTGTGAGTCTAATAAAGAAGGTTAAAATCCCTCTGCCTGCCG 150
             || ||  |||||||||||||||||| ||||| || | ||||||||||||  || |   |||
Sbjct: 12646 TTAATTTAAAACATTAGATTGTGATTCTAAAAATGGAGGTTAAAATCCTCCTACTCACCG 12705

Query:   149 AGAGGGGC 142
             |||  |||
Sbjct: 12706 AGAAAGGC 12713
```

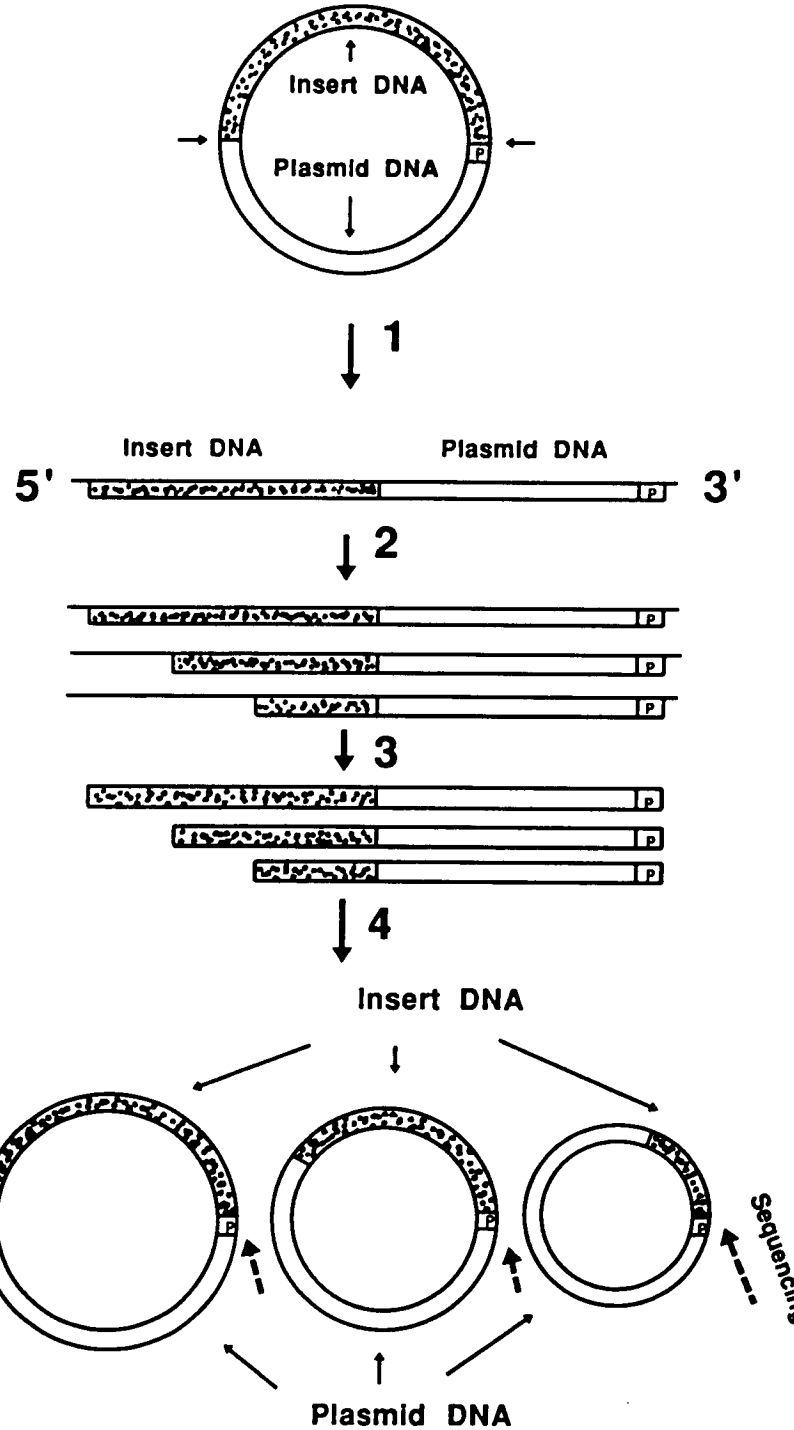
SEQUENCING OF LONG DNA FRAGMENTS

The DNA fragments inserted into the plasmid were 3-6 kb, which were too long to sequence conveniently from a single primer site on the vector. A number of methods have been developed for sequencing a large DNA fragment. A method called nested sets of deletions (Henikoff, 1984) was used to sequence the four large fragments. The fragments were deleted unidirectionally from the 5'-end, generating predetermined lengths of DNA to perform subcloning (Fig 2-8). Overlapping clones were sequenced using an automated DNA sequencer.

Nested Deletions and Subcloning

A 500 ml culture of each recombinant clone provided 30-80 µg of supercoiled DNA, with which nested sets of deletions were obtained. Exonuclease III, S1 nuclease, Klenow DNA polymerase, T4 DNA ligase, and buffers were purchased from Promega. *BamH I* and *Xho I* generated 5' overhangs and *Kpn I* and *BstX I* made 3' overhangs that protected the priming site from the activity of Exonuclease III. Each digestion was carried out according to the manufacturer's protocol and completion of restriction enzyme activity was checked by running a small part of the reaction on a 1% agarose gel. The digested DNAs were purified with phenol/chloroform and precipitated with ethanol. 5-7 µg of DNA digested with two enzymes was dissolved

Figure 2-8. Schematic diagram of nested deletions. The supercoiled recombinant DNA was prepared by a maxiprep method. The shaded part represents the insert DNA (lamprey mtDNA) and P indicates the priming site of the cloning vector. 1. The clone DNA is linearized by two restriction enzymes generating 5' and 3' overhangs. The priming site is protected by the 3' overhang from the *Exo III* activity in the following step. 2. The linearized DNA was digested from the 5' overhang with *Exo III* and DNA aliquots were removed in every 30 seconds. 3. The single strand was got rid of by S1 nuclease. 4. Using Klenow, dNTPs, and DNA ligase, the timed aliquots were religated. After transformation, plating, and size selections, the clones with appropriate sizes were sequenced.



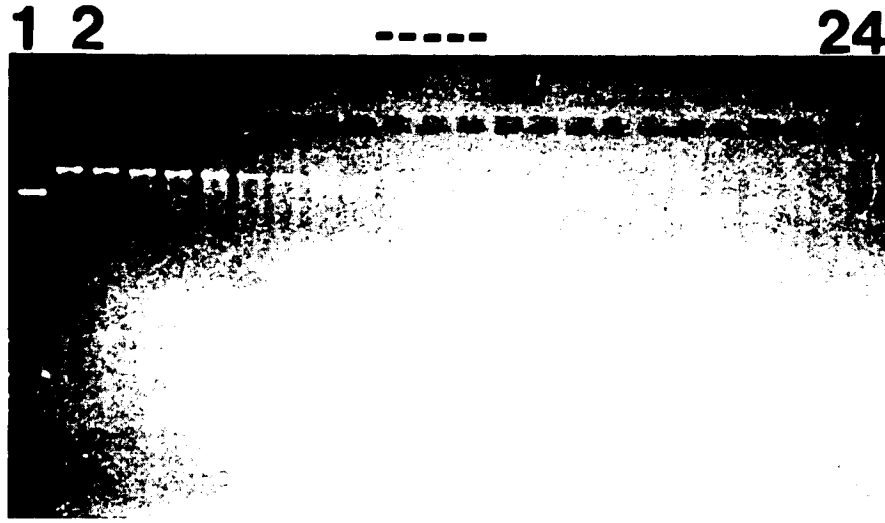
in 1x Exonuclease III buffer. The number of tubes needed for sets of deletions were estimated from the length of the DNA fragments. For example, I wanted to get a 250 base difference in length between two consecutive deletions. Thus for 6 kb fragment, 6,000 bp ÷ 250 bp = 24 deletions needed. 7.5 µl of S1 nuclease mix [for 25 deletions, 172 µl of dH_2O, 27 µl of S1 7.4x buffer (0.3 M potassium acetate, pH4.6, 2.5 M NaCl, 10 mM $ZnSO_4$, and 50% glycerol), and 60 units of S1 nuclease] were added to each of 24 tubes and placed on ice. After the DNA was warmed up in a 37°C water bath for 5 min, 500 units of Exo III was added and mixed as rapidly as possible. 2.5 µl aliquots of DNA were removed and added to the S1 mix tubes on ice, at 20 second intervals. After all the DNA was taken, the S1 tubes were moved to room temperature for 30 min. 1 µl of S1 stop buffer (0.3 M Tris-base/0.05 M EDTA) was added to the tubes and incubated at 65°C for 10 min to inactivate S1 nuclease. 2 µl from each time point were run in a 1% agarose gel to check the extent of digestion. When the digestion worked properly, the sizes of DNA decrease linearly among the time points.

After all time points were transferred to room temperature from 65°C, 1 µl of fresh Klenow mix [30 µl of Klenow buffer (20 mM Tris-HCl, pH 8.0/100 mM $MgCl_2$), and 5 units of Klenow DNA polymerase] was added to each tube. After a 3 min incubation, 1 µl of dNTP mix (0.125 mM of dNTP) was added and incubated an additional 5 min at 37°C. The samples were transferred to room temperature and 40 µl of ligase mix [790 µl of dH_2O, 100 µl of 10x ligase buffer (500 mM Tris-HCl, pH 7.6, 100 mM $MgCl_2$, and 10 mM ATP), 100 µl 50% PEG, 10 µl 100 mM DTT, and

5 units of T4 DNA ligase] were added to each sample and incubated for 1-3 hours at room temperature. 10 µl of the ligated DNA from each time point were plated as described before. The differences from the original cloning were that no agents for color selection such as X-gal and IPTG were added to the LB agar because all colonies were supposedly white in color.

Theoretically, all clones from the same time point contained an identical size. However the sizes of clones recovered from each time point varied widely. Consequently, a screening before sequencing was necessary (Barnes, 1977). I divided a petri dish containing LB agar and two antibiotics into 16 sections with a marker and streaked bacterial colonies from a time point on the surface of each section. After incubation at 37°C overnight, a portion of bacteria grown from each section was taken in 16 tubes. 50 µl of 10 mM EDTA, pH 8.0 and lysis buffer (for 50 ml, 2 ml 5M NaOH, 2.5 ml 10% SDS and 10g sucrose) were added to each tube respectively and mixed thoroughly. After a 5 min incubation at room temperature, 1.5 µl of 4 M KCl and 0.5 µl of 0.4% bromophenol blue were added and vortexed. The tubes were placed on ice for 5 min and centrifuged at 12,000 x g for 3 min at 4°C. 10 µl of the supernatant was loaded on a 1% agarose gel. The clones showing a series of sizes were selected and miniprepped. The final DNA was checked on a 1% agarose gel again before sequencing in order to see the sizes of subclones and to quantify the DNA before sequencing (Fig. 2-9).

Figure 2-9. Agarose gel (1%) picture of timed aliquots of the nested deletions. The first lane shows the undigested (circular) recombinant DNA. The sizes of DNA from each time point were linearly reduced approximately 300 bp in length (lanes 2 through 24).



DNA Sequencing

All nucleotide sequences were obtained from double-stranded recombinant DNA. Prepared plasmid DNA from either the standard alkaline lysis miniprep (Sambrook et al., 1989) or the Magic miniprep kit (promega) was used for Taq DyeDeoxy Terminator cycle sequencing reactions (Applied Biosystems Inc.). For each reaction, 4 µl 5x TACS buffer, 1 µl dNTP mix, 1 µl of each DyeDeoxy terminators, 3 µl of M13 primer, and about 1-2 µg of DNA template were added. 96°C (30 sec), 50°C (15 sec), and 60°C (4 min) were used for denaturation, annealing, and extension respectively. After 25 cycles in Perkin-Elmer Cetus thermal cycler, the unincorporated dye terminators were removed from the cycle sequencing reaction with a spin column containing 5% sephadex by centrifugation at 1,800 x g for 2 min. Finally the entire product of cycle sequencing was resuspended in 4 µl of formamide-EDTA buffer, denatured, and loaded on an automated DNA sequencer (ABI 373A). The sequences from subclones overlapped as predicted.

CONCLUSION

Fresh samples from highly active tissues are the best source for intact mtDNA isolation. As a consequence of damage, in most cases, slowly frozen tissues are not adequate for isolation of supercoiled mtDNA. The use of a tissumizer for homogenization and ultra-centrifugation in CsCl-ethidium bromide contribute greatly to the yield of intact and pure mtDNA. Other methods seem to be much more tricky to follow. Actually the phenol/chloroform extraction method was not effective for isolation of intact mtDNA because it could not completely remove linear DNA, causing too much background. In addition, the speed of low centrifugations is one of the critical factors. Thus the speed should be adjusted by comparisons of the amount of mtDNA obtained from a series of different speeds.

Preparing frozen stocks of competent cells and clones is highly recommended, since it saves time and labor without a significant difference in results. Non-radioactive labelling is as sensitive as the radioactive counterpart for Southern blotting, and much easier to handle. It seems to be obvious that the BLAST search is a powerful identification tool if other methods are limited. The nested deletion method is highly effective for sequencing large DNA fragments under circumstances where the facilities are available to sequence double-stranded DNA. Needless to say, an automated DNA sequencer accelerates the progress of sequencing if the DNA is appropriately prepared.

CHAPTER III

COMPLETE SEQUENCE OF THE SEA LAMPREY (PETROMYZON MARINUS) MITOCHONDRIAL GENOME: A UNIQUE GENE ORDER AND SIGNIFICANCE FOR THE EVOLUTION OF MITOCHONDRIAL GENOME STRUCTURE

ABSTRACT

The complete nucleotide sequence of the sea lamprey (*Petromyzon marinus*) mitochondrial genome has been determined. The lamprey mt genome is 16,201 bp in length, and contains genes for 13 proteins, 2 rRNAs, 22 tRNAs and 2 major non-coding regions. The order and transcriptional polarities of protein-coding genes are basically identical to those of other chordate mtDNAs, demonstrating that the common mitochondrial gene organization of vertebrates established at an early stage of vertebrate evolution. The major noncoding region is separated by two tRNA genes. The first region probably functions as the control region, because it contains distinctive conserved sequence blocks (CSB-II and III) common to other vertebrate control regions. The central conserved domain observed in other vertebrate control regions is not found in the lamprey, suggesting that it is a recently evolved functional domain. Non-coding segments are not found in the expected position of the O_2, which suggests either that one of the tRNA genes has a dual function, or that the second non-coding region may function as O_2. The base composition at the wobble positions of four-fold

degenerate codon families is highly biased toward thymine (32.7%). Values of GC- and AT-skew are typical of vertebrate mitochondrial genomes.

INTRODUCTION

Lampreys, among the earliest diverged vertebrates, have a unique evolutionary history in the 550 million years since they separated from the main vertebrate lineage. Approximately 40 species are distributed over coastal drainages throughout the Northern Hemisphere. Four species are found in temperate areas of the Southern Hemisphere (Moyle et al., 1988). They have two distinct life history patterns. Before migrating to the ocean or lakes, sea lampreys spend most of their life time as ammocoetes and spend up to 2 years as adults feeding on fishes. On the other hand, fresh water lampreys spend their whole life in shallow streams, without migration to the ocean or lakes. The fresh water lampreys are not parasitic. Mostly because of the lack of distinct characters, the taxonomy of lamprey species is not well established (Hubbs, 1971). As one of the earliest diverged vertebrates, it attracts particular attention from evolutionists, because it is expected that the lamprey may provide useful information about the evolutionary history of vertebrates, especially relationships between the early diverged vertebrates (Stock et al., 1992; Forey et al., 1993).

The patterns of gene arrangements among animal mitochondrial genomes may be informative for the phylogeny of distantly related taxa, because the rate of gene rearrangement is much slower than nucleotide substitutions (Brown, 1985; Smith et

al., 1993). The sequential gene order within each phylum shows a basic arrangement despite minor relocations of genes in some sublineages (Hoffmann et al., 1992). In nematodes, the gene orders of two species (*C. elegans* and *A. suum*) display only the transposition of the A+T rich region (Okimoto, et al., 1992). In insects, 11 tRNA genes have been moved between the genomes of honey bee and *D. yakuba* but most keep the same transcriptional polarity (Garesse, 1988; Crozier et al., 1993). Exceptions to the conservation of gene order within phyla are found in echinoderms. Partial sequences demonstrate that brittle stars and sea stars share a nearly identical order, which is different from that of sea urchin and sea cucumber (Smith et al; 1993). The change of gene order can be explained by a simple inversion of a 4.6 kb segment. Vertebrate mtDNAs also show highly conserved gene order from bony fish to human, although minor rearrangements have been found in marsupial and bird mtDNAs (Desjardins, 1990; Janke et al., 1994). The bird genome shows a transposition of genes for ND6 and tRNA-Glu relative to the genes for Cyt b, tRNA-Thr, and tRNA-Pro. In the marsupial genome, five tRNA genes are relocated within the cluster of tRNA genes near the L-strand replication origin.

As more sequence data accumulate, the control region appears to be the most enigmatic and the least understood portion of the animal mitochondrial genome. Although the functions of the control region of mammalian mtDNA have been reported (Clayton, 1982), many questions about the function and evolution of the region still remain unanswered. In particular the function of the central conserved

region found in vertebrates, but not other deuterostomes, is not known.

In this chapter, I present the complete nucleotide sequence of the sea lamprey mitochondrial genome. Comparisons of lamprey mt genome with other vertebrate and invertebrate genomes may yield insight into the mechanisms responsible for the structural evolution and patterns of substitutions in animal mitochondrial genomes. These data may also be useful to elucidate the phylogeny of early vertebrates and will facilitate population studies of the widely distributed lamprey species.

MATERIALS and METHODS

MtDNA Isolation and Cloning

Adult sea lampreys were collected from fish ladders on the Cocheco River at Dover, NH during the spawning season of 1992 (from May to June). Pure mitochondrial DNA was obtained from the fresh liver and eggs using slight modifications of the protocols described by Lansman et al. (1981) and Dowling et al. (1990). The low speed centrifugations were performed at 550 x g instead of 700-800 x g and ultracentrifugation of CsCl-Ethidium bromide gradient at the density of 1.50 g/ml, instead of 1.55 g/ml. The isolated mtDNA was digested with *EcoRI*, resulting in four fragments (3, 3.5, 4, 6 kb) covering the whole genome.

Each of the *EcoRI* fragments was cloned in pBluescriptII SK and amplified using *E.coli* XL1-Blue (Stratagene). Nested sets of deletions were constructed from the four recombinant DNAs with the Erase-A-Base system kit (Promega). After checking the sizes of the inserts on a 1.0% agarose gel, overlapping clones were prepared and sequenced as follows. For detail, see chapter II.

DNA Sequencing

All nucleotide sequences were obtained from double-stranded plasmid DNA. Plasmid DNA from either standard alkaline lysis miniprep (Sambrook et al., 1989) or

Magic miniprep kit (Promega) was used for Taq DyeDeoxy Terminator cycle sequencing reaction (Applied Biosystems Inc.). Extra dye terminators were removed from the cycle sequencing reaction with a spin column containing 5% Sephadex. Finally the entire product of cycle sequencing was dried and resuspended in 4 µl of formamide-EDTA buffer, denatured at 90°C, and loaded on an automated DNA sequencer (ABI 373A). The sequence obtained from each subclone averaged 350 bp, and overlapped the next clone for 100-150 bp. There was no sequence variation observed within the overlapping regions in any clones.

After determining the entire nucleotide sequence, I designed 4 pairs of primers on the ends of 4 fragments with help of a computer program (Lasergene, DNAStar Inc.) in order to see if there were any missing *EcoRI* fragments. The PCR products were about 300-400 bp in length, overlapping the junctions of two neighboring fragments. The sequences of the PCR products confirmed that the four *EcoRI* generated fragments covered the entire sea lamprey mitochondrial genome.

Sequence Analyses

The sequences of subcloned plasmids were aligned with SeqEd (ABI), and the entire nucleotide sequence was further analyzed with ESEE (Cabot, 1988) and GCG (Genetic Computer Group, Version 7.0). The 13 protein-coding genes were determined by comparisons of DNA or amino acid sequences of other mitochondrial genomes. The 22 tRNA genes and 2 rRNA genes were also identified by sequence homology,

secondary structure and/or anticodon sequences. The rRNA sequences were compared by COMPARE and DOTPLOT, the secondary structures were folded by FOLD and SQUIGGLES, and the comparisons of multiple sequences were made with PILEUP programs in GCG package.

The base composition and codon frequency were obtained with COMPOSITION and CODONFREQUENCY programs in GCG. Base frequency statistics were calculated with the formulas in Perna and Kocher (1994): %GC = proportion of G + C out of the total base number, GC-Skew = (G-C)/(G+C), and AT-Skew= (A-T)/(A+T)

RESULTS and DISCUSSION

Genome Content

The gene content of the sea lamprey mitochondrial genome includes 13 proteins, 22 tRNAs, 2 rRNAs, and 2 major non-coding regions (Fig. 3-1). As seen in other vertebrate mtDNAs, most genes are encoded on the first, or heavy-strand, except for ND6 and 8 tRNA genes. The sequence of all nucleotides of the first strand is presented in Figure 3-2 and the locations of the genes are shown in Table 3-1. The length of the lamprey mitochondrial genome is the shortest yet seen in vertebrates because of the small size of the putative control region. The sizes of most of the other genes are similar to other vertebrate mtDNAs. One exception is observed in Cyt b gene. Lamprey Cyt b gene is the longest among animal mtDNAs sequenced so far (Table 2), approximately 50 bp (16 amino acids) longer than other vertebrate Cyt b genes.

There are two unassigned DNA segments between the ND6 and Cyt b genes. The segments are divided by two tRNA genes, Thr and Glu. One of the two major non-coding regions (the first part) likely represents the control region containing the regulatory sequences for the replication origin of the heavy strand. There are two sets of tandem repeats found in the two major non-coding regions. The first region contains a tandem repeat of 3 x 39 bp and the second has a 7 x 26 bp repeat

Figure 3-1. The genetic map of sea lamprey mitochondria. All genes are drawn to the scale of sizes. Protein-coding genes and two major non-coding regions are shown in side of boxes. tRNA genes are represented by 1-letter amino acid outside or inside circles according to the coding strand. All outside tRNA genes are encoded on the first strand with clockwise transcriptional polarity (filled arrowhead). All protein-coding genes except for ND6 are the first-strand encoded with clockwise transcription polarity (empty arrowhead). The 'rpt' stands for repetitive sequences in the second major non-coding region. Four *EcoRI* restriction sites are shown, which were used for cloning and sequencing. All abbreviations are given in the text.

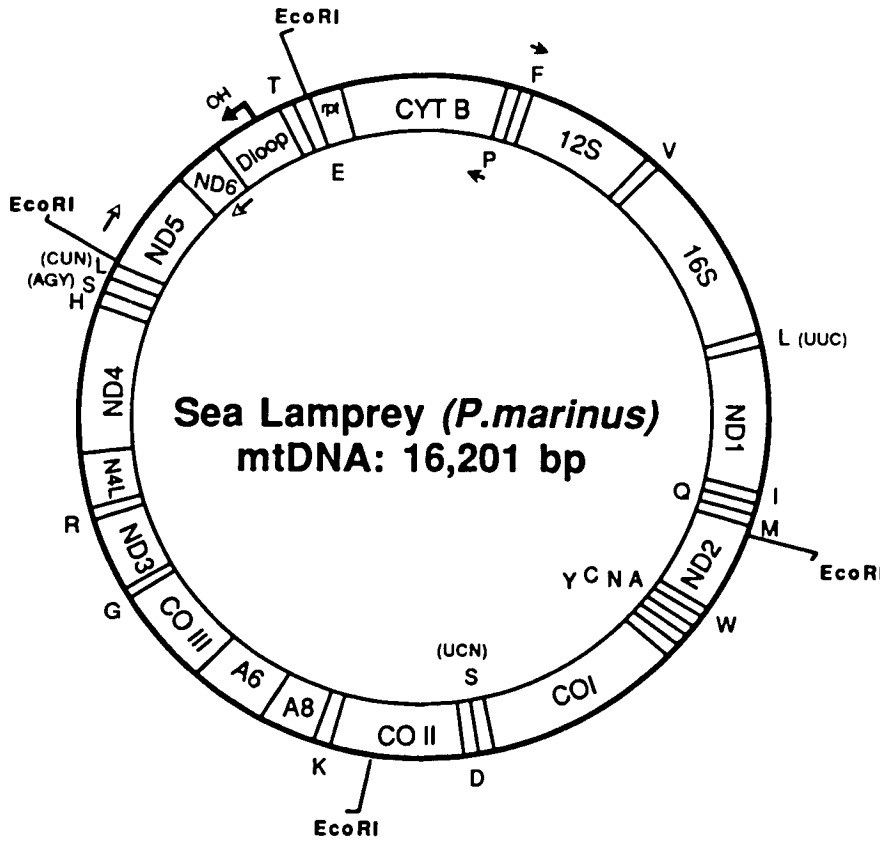
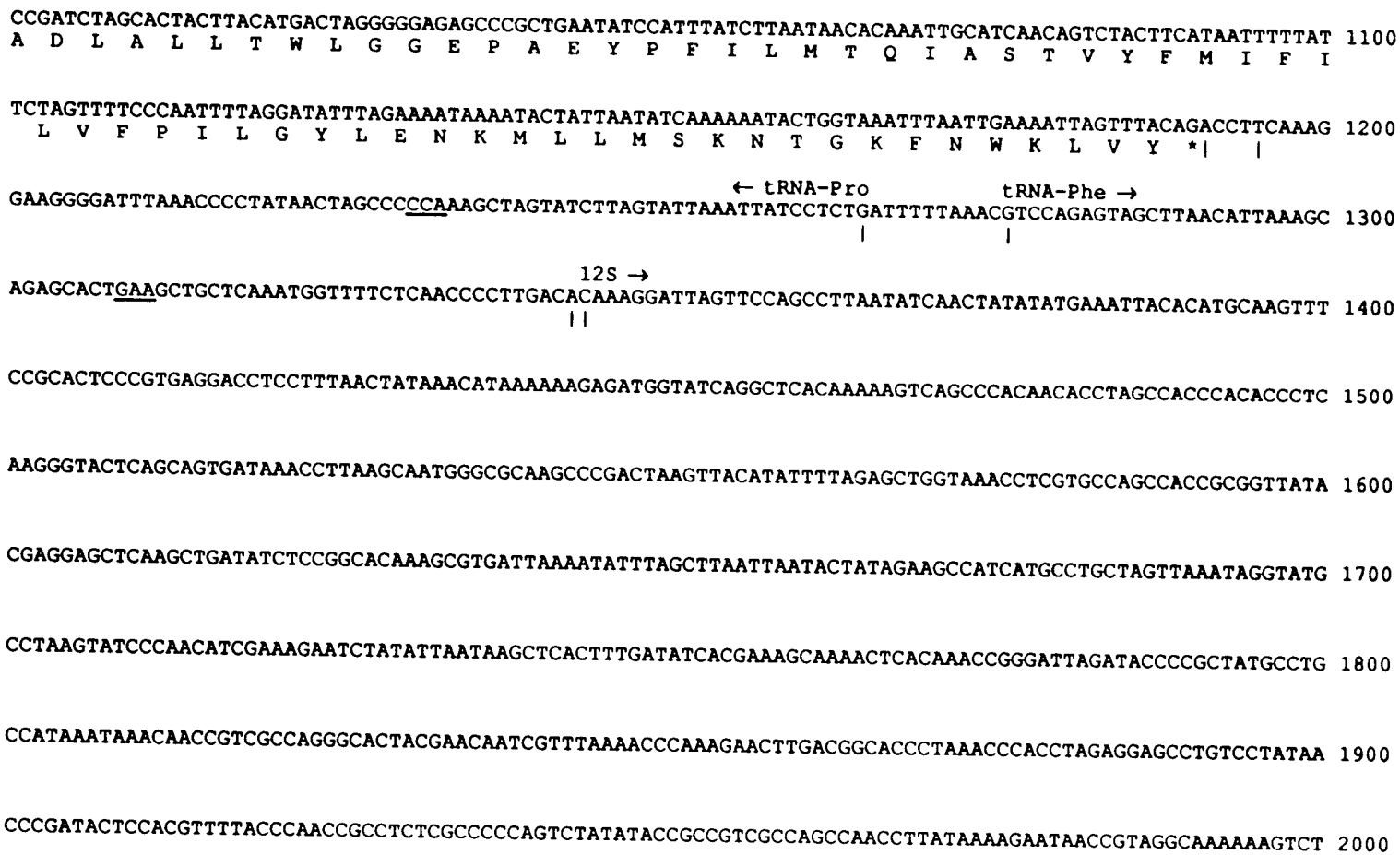


Table 3-1. Location and coding strand of each gene in the sea lamprey mitochondrial DNA. The asterisks indicate gene overlaps with the following gene.

Name of gene	Location	Strand
Cyt b	1-1191	First
tRNA-Pro	1195-1265	Second
tRNA-Phe	1276-1343	First
12S rRNA	1344-2243	First
tRNA-Val	2244-2314	First
16S rRNA	2315-3935	First
tRNA-Leu (UUR)	3936-4009	First
ND1	4014-4979	First
tRNA-Ile	5003-5071	First
tRNA-Gln	5074-5144	Second
tRNA-Met	5145-5212	First
ND2 *	5214-6257	First
tRNA-Trp	6256-6323	First
tRNA-Ala	6325-6393	Second
tRNA-Asn	6397-6465	Second
tRNA-Cys	6467-6532	Second
tRNA-Tyr	6538-6608	Second
COI *	6610-8163	First
tRNA-Ser (UCN)	8154-8224	Second
tRNA-Asp	8226-8294	First
COII	8298-8987	First
tRNA-Lys	8996-9061	First
ATP8 *	9064-9231	First
ATP6 *	9222-9935	First
COIII	9901-10686	First
tRNA-Gly	10695-10764	First
ND3	10765-11115	First
tRNA-Arg	11120-11185	First
ND4L *	11194-11484	First
ND4	11478-12854	First
tRNA-His	12856-12924	First
tRNA-Ser (AGY)	12925-12994	First
tRNA-Leu (CUN)	12995-13066	First
ND5 *	13068-14864	First
ND6	14849-15367	Second
D-loop	15368-15858	First
tRNA-Thr	15859-15930	First
tRNA-Glu	15932-16002	Second
Repeats	16003-16201	First

Figure 3-2. Complete sequence of lamprey mitochondrial genome. The sequence is from the first strand and the transcriptional directions are marked by arrows (→). The vertical lines (|) indicate the beginning and ending positions of each gene. The amino acid translations presented below the nucleotide sequence used the mammalian mitochondrial genetic code and stop codons are designated by asterisks (*). The numbers are started from the first nucleotide of Cyt b. The abbreviations are described in text.

(Figure 3-2 continued)



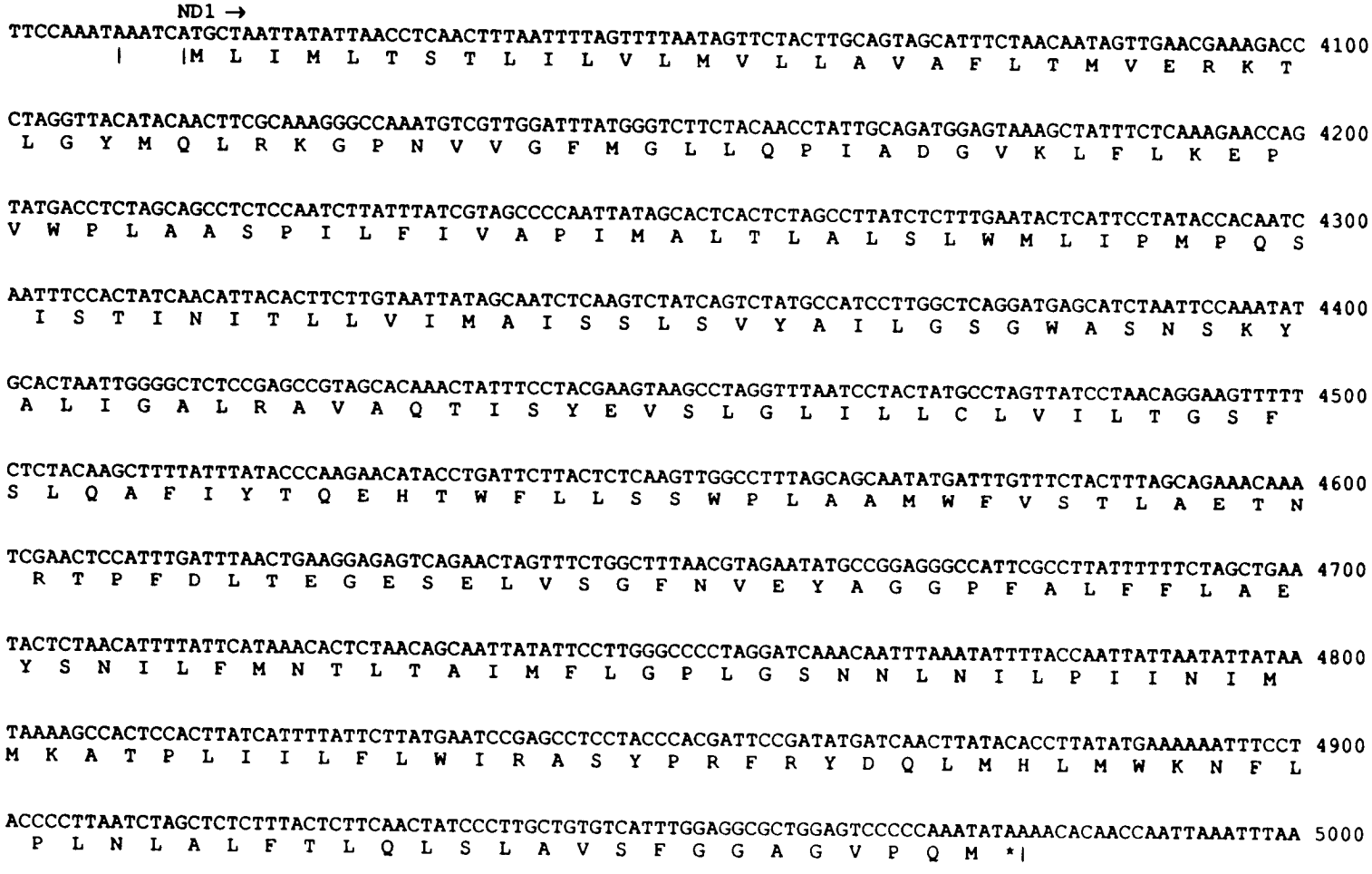
(Figure 3-2 continued)

```
ATCTATACAAATACGTCAGGTCGAGGTGCAACCTATGAGGCAGGCAGAGATGGGCTACACTCTCTACCCAGAGTATACGAATAATTTAATGAAAAAATTT 2100

TGAAGGTGGATTTAGCAGTAAACAAGAATAGTTTGTCTAGTTGAAGTTGCCACTAGGGTGCGTACACACCGCCCGTCACTCTCCCCCCACACCCGGAGAA 2200

                                                tRNA-Val →
AAGTCGTAACATGGTAAGCGTACCGGAAGGTGCGCTTGGAAAACAGAAGATAGCTTAAAAGTTAAGCATTTCCCTTACACCGAAAATATTCTTGTGCAAT 2300
                                                ||

                 16S →
TCAAGATCTTCTGACTACTGATCTAAAGATATATTTCTAACAACTCTTAACTTCTGATTATAAAACAATTAAATACTTTACCGCAAACCATTGCCCCCAT 2400
                 ||

TTTAGTATAGGTGATAGAAAAAAAATTATACACATAATAGTACCGCAAGGGAATATTGAAAAAGAAGTGAAATAAATTGATTAAGTAAAACAAAGCAAAG 2500

ATTAAATCTTGTACCTTTTGCATCATGGCTTAGCAAGCAAACCCGAATATACTGCCGCCACCCCGAAACTAGACGAGCTACCCTGGGATTACCTATAAGG 2600

GTTAATCCGTATCTGTGGCAAAAGATTGGAAAAACCCCTGGGTAGAGGTGAAAAGCCTACCGAGCCTAGTGATAGCTGGTTACTTAAGAAACAAGTTTAA 2700

GCTTGATCTTAACTTGTAGATGAGCAACAAAATTACTTAGAAATTTAAACATCTTACTCCTCTACACTTTAAGTTTTATTCACTAGGGGTACAGCCCTAG 2800

TGAACAGAGATACAGCTCTATTAATTAGATAATATACCACATTTTAAACTTAAGTAGGCCTAAAAGCAGCCACCAAGAAGAAAAGCGTTACAGCTTAAGT 2900

TTACTAAATTATAAATACCAAAATATATAAAGACCCTATAAACACTATTAAGTAATCCTATAAAATAGGAGATATCCTGCTAAGATTAGTAATTTGAGCC 3000
```

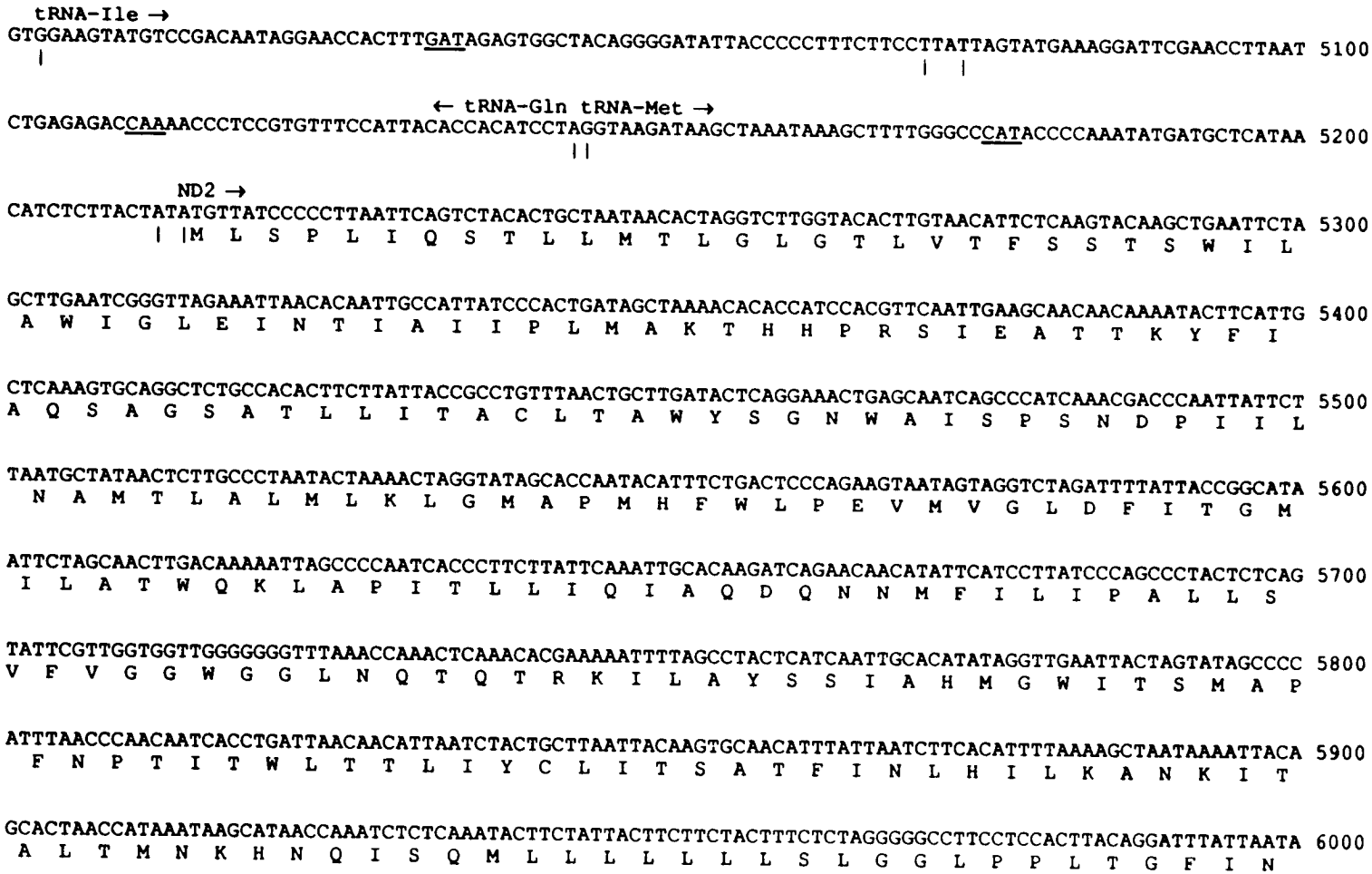
(Figure 3-2 continued)

```
CGCACCCCTCTAAATGTAAGTGTACACCAGATCGGACCAACCACTGGAAATTAACGGCCCTTAAAACAACAGGAAGTCAGAAACATAAACAACAACAAGA 3100

AAAACAAGAACTAATAACCGTTAACCCTACACTGGAACATAATATAGAAGATATAAAGGATAAGAAGGAACTCGGCAAACACATGCCTCGCCTGTTTACC 3200

AAAAACATCACCTCCAGATAAAAATCAAGTATTGGAGGCAAGACCTGCCCAATGATTAATATTGAATGGCCGCGGTACTTTGACCGTGTAAAAGTAGCGT 3300

AATCACTTGTCTTGTAAATTAAGACTGGAATGAAAGGTTACACGAGGGCATAACTGTCTCCTTATCCCTATCAATGAAATTGACCTACCCGTGCAAAGGC 3400

GGGTATAAACCCATAAGACGAGAAGACCCTGTGGAGCTTCCAAACATTTACATCGAATAATAATTATTTACGATGTACAGTTTTAGGTTGGGGCAACCAC 3500

GGAACAAAAGTAATATCCACGACGACGAAAATATAATTTTCTAAGCCTAGAACCACAACTCTAAGCACTAGTAAAACTAACGTTAATAGACCCAGCATCA 3600

CTTGCTGACTAACGAAACAAGTTACCCCAGGGATAACAGCGCAATCCTTTCCACGAGCCCGAATCAACGAAAGGGTTTACGACCTCGATGTTGGATCGGG 3700

GCACCCCAATGGCGCAAAAGCTATTAAAGGTTCGTTTGTTCAACGATTAAAGCCCCACGTGATCTGAGTTCAGACCGGAGTAATCCAGGTCAGTTTCTAT 3800

CTATGTTTGCTGTTTCCCTAGTACGAAAGGACCGGTGAAACAAGGGTCTATACACTTATGCAAACCCTACATCAATCCTATGAAACCAACTCAATAAGAA 3900

                                    tRNA-Leu(UUR) →
TAGTAAGCAACATTAAATAATAACTAGATAAGTTTATTTGGATGGCAGAGTTCAGTAATTGCACAAGGTTTAAGCCCTTATACCAGAGGTGCAAATCCTC 4000
                                    ||
```

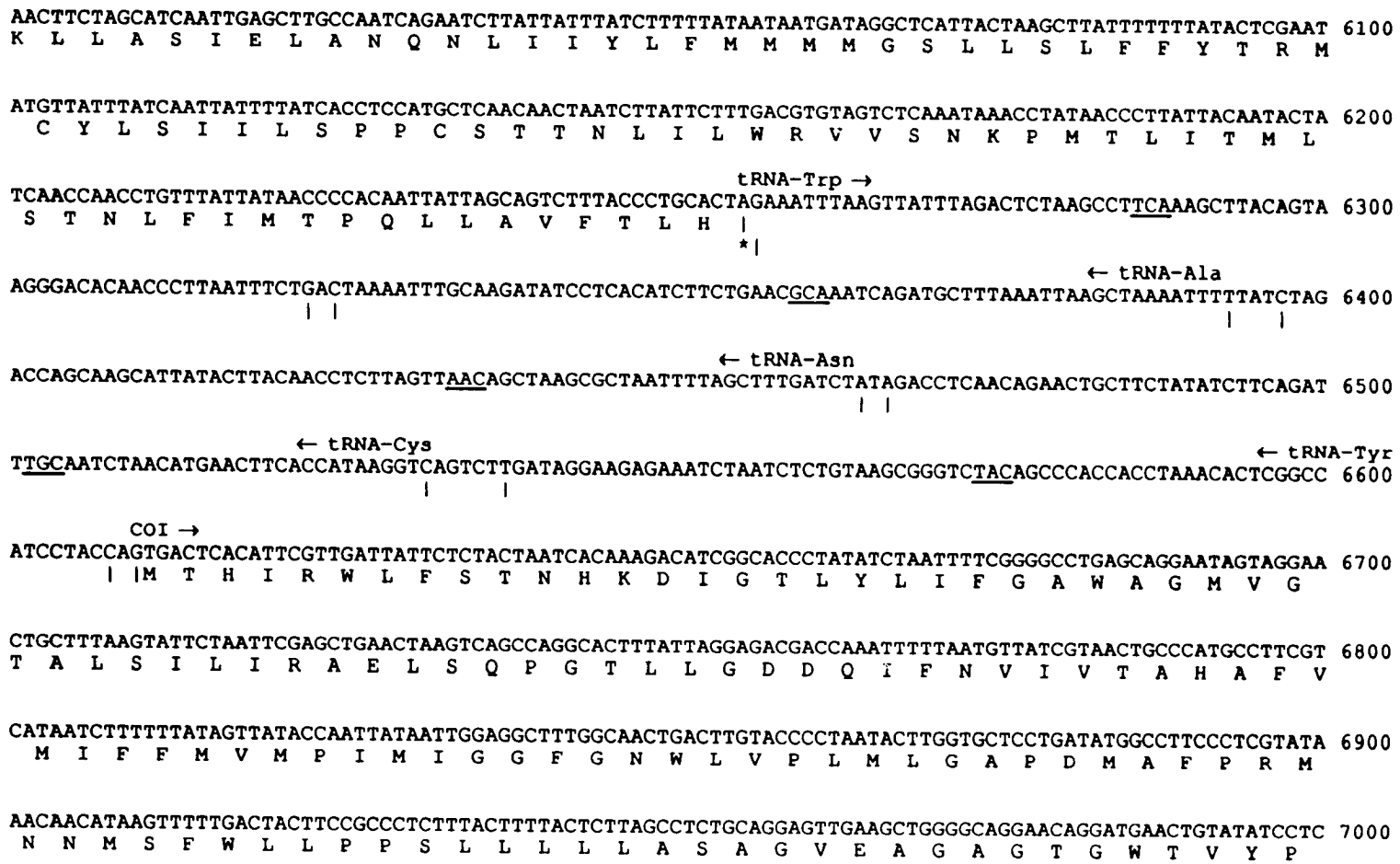
(Figure 3-2 continued)



(Figure 3-2 continued)



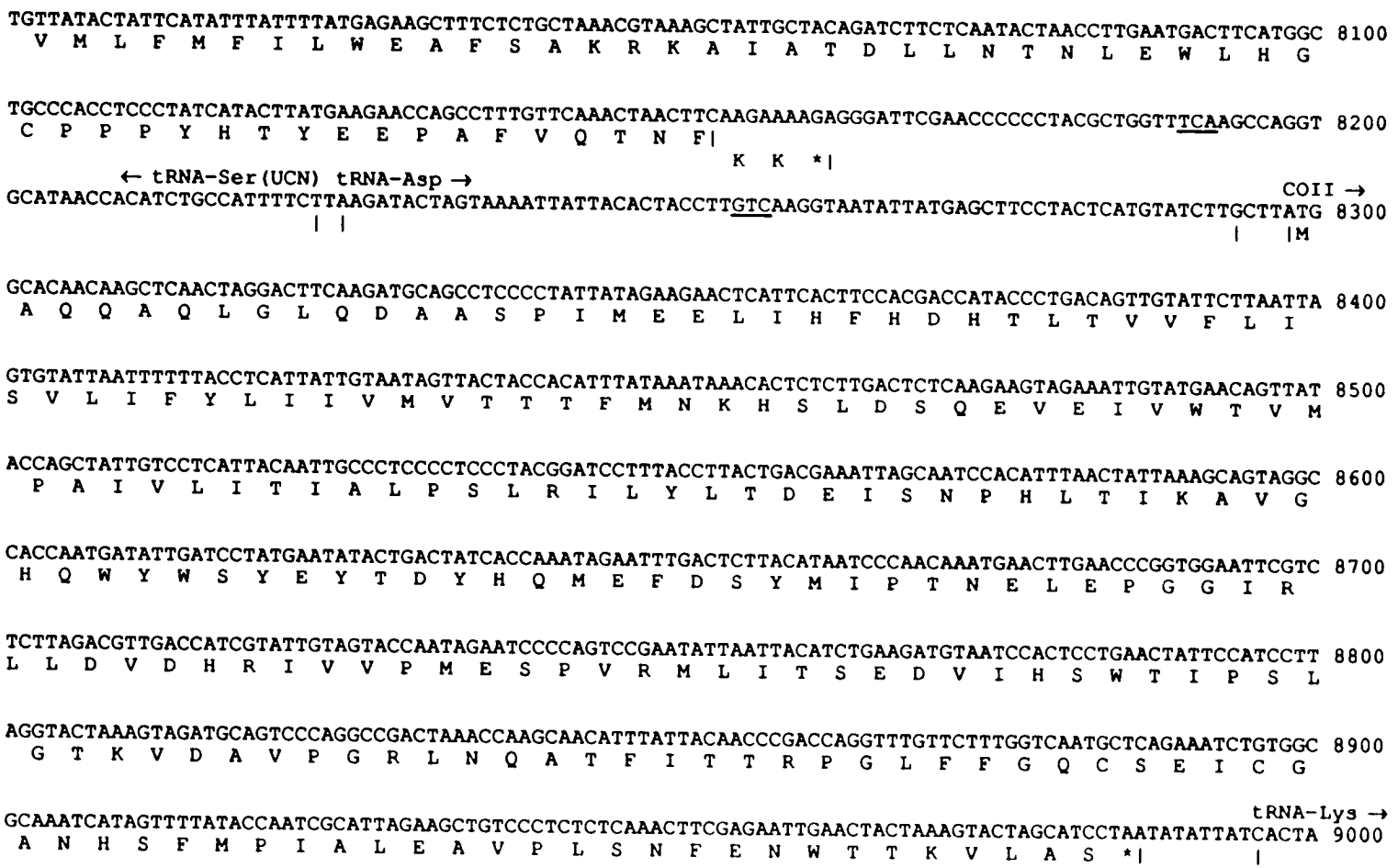
(Figure 3-2 continued)



(Figure 3-2 continued)

```
CCTTAGCCGGAAACCTAGCCCACACCGGGGCCTCTGTCGACCTAACAATCTTTTCCTTACACTTAGCCGGAGTTTCATCAATTCTAGGAGCAGTTAATTT 7100
P  L  A  G  N  L  A  H  T  G  A  S  V  D  L  T  I  F  S  L  H  L  A  G  V  S  S  I  L  G  A  V  N  F

CATCACAACTATTTTTAACATGAAACCCCCAACTATGACTCAATaCCAAACCCCCTTATTTGTTTGATCAGTCTTAATCACTGCAGTTCTTCTTCTTCTA 7200
   I  T  T  I  F  N  M  K  P  P  T  M  T  Q  Y  Q  T  P  L  F  V  W  S  V  L  I  T  A  V  L  L  L  L

TCTCTACCAGTACTAGCAGCTGCTATCACAATACTTCTAACAGATCGTAACTTAAATACATCCTTCTTCGACCCTGCAGGAGGAGGAGACCCCATTCTTT 7300
   S  L  P  V  L  A  A  A  I  T  M  L  L  T  D  R  N  L  N  T  S  F  F  D  P  A  G  G  G  D  P  I  L

ACCAACACTTATTTTGATTCTTCGGACACCCTGAAGTTTATATTCTAATTCTTCCAGGCTTCGGAATTATTTCACACGTAGTTGCTTATTATGCTGGGAA 7400
Y  Q  H  L  F  W  F  F  G  H  P  E  V  Y  I  L  I  L  P  G  F  G  I  I  S  H  V  V  A  Y  Y  A  G  K

AAAAGAACCATTCGGATATATAGGAATAGTTTGAGCAATAATAGCCATTGGACTACTAGGATTTATTGTTTGAGCTCATCACATATTTACAGTAGGAATA 7500
   K  E  P  F  G  Y  M  G  M  V  W  A  M  M  A  I  G  L  L  G  F  I  V  W  A  H  H  M  F  T  V  G  M

GACGTTGATACACGAGCCTATTTTACATCAGCCACAATAATTATTGCTATCCCAACAGGAGTCAAAGTCTTCAGTTGATTAGCCACTCTTCATGGAGGAA 7600
   D  V  D  T  R  A  Y  F  T  S  A  T  M  I  I  A  I  P  T  G  V  K  V  F  S  W  L  A  T  L  H  G  G

AAATCGTATGACATACCCCTATATTATGAGCCCTAGGTTTTATTTTCTTATTTACTGTAGGAGGACTCACAGGAATTGTTTTATCAAATTCATCACTAGA 7700
K  I  V  W  H  T  P  M  L  W  A  L  G  F  I  F  L  F  T  V  G  G  L  T  G  I  V  L  S  N  S  S  L  D

CATTATTCTTCATGACACTTACTATGTTGTAGCCCATTTCCATTATGTTCTATCTATAGGAGCTGTTTTCGCAATTATAGCAGGATTTGTCCACTGATTC 7800
   I  I  L  H  D  T  Y  Y  V  V  A  H  F  H  Y  V  L  S  M  G  A  V  F  A  I  M  A  G  F  V  H  W  F

CCACTATTTACAGGATATACACTTAACGAAACCTGAGCAAAAGCTCATTTCATTATTATGTTTGCTGGTGTTAATCTTACATTCTTCCCTCAACACTTCC 7900
P  L  F  T  G  Y  T  L  N  E  T  W  A  K  A  H  F  I  I  M  F  A  G  V  N  L  T  F  F  P  Q  H  F

TAGGTCTAGCTGGAATACCACGACGTTACTCAGACTACCCAGATGCTTATACTACATGAAATATTATTTCCTCAATTGGGTCAACAGTCTCACTAATCGC 8000
L  G  L  A  G  M  P  R  R  Y  S  D  Y  P  D  A  Y  T  T  W  N  I  I  S  S  I  G  S  T  V  S  L  I  A
```

(Figure 3-2 continued)



(Figure 3-2 continued)

```
                                                                                ATP8 →
AGAAGCTAACTTAGCATCAGCCTTTAAGCTGAAGATGGGCGAATACCTTCTCCCTTAGTGATATGCCACAACTCGATCCTGCCCCTTGATTCTCTATAC 9100
                                                                                 |  |M  P  Q  L  D  P  A  P  W  F  S  M

TTACAGTATCATGACTAATTATTTTTCTCTTAATTATACCAACTATCTTATTTTATCAACCACAAAACACCATCTCTACTAAACAAGTTACTAAACCCAA 9200
L  T  V  S  W  L  I  I  F  L  L  I  M  P  T  I  L  F  Y  Q  P  Q  N  T  I  S  T  K  Q  V  T  K  P  K

                        ATP6 →
ACAATCCACTTGAACCTGACCATGACACTAGATATCTTTGACCAATTTACCTCCCCAACAATATTTGGGCTTCCACTAGCCTGATTAGCTATACTAGCCC 9300
  Q  S  T  W  T  W  P|M  T  L  D  I  F  D  Q  F  T  S  P  T  M  F  G  L  P  L  A  W  L  A  M  L  A
                        W  H  *|

CTAGCTTAATATTAGTTTCACAAACACCAAAATTTATCAAATCTCGTTATCACACACTACTTACACCCATCTTAACATCTATTGCCAAACAACTCTTTCT 9400
P  S  L  M  L  V  S  Q  T  P  K  F  I  K  S  R  Y  H  T  L  L  T  P  I  L  T  S  I  A  K  Q  L  F  L

TCCAATAAACCAACAAGGGCATAAATGAGCCTTAATTTGTATAGCCTCTATAATATTTATCTTAATAATTAATCTTTTAGGATTATTACCATATACTTAT 9500
  P  M  N  Q  Q  G  H  K  W  A  L  I  C  M  A  S  M  M  F  I  L  M  I  N  L  L  G  L  L  P  Y  T  Y

ACACCAACTACCCAATTATCAATAAACATAGGATTAGCAGTGCCACTATGACTAGCTACTGTCCTCATTGGGTTACAAAAAAAACCAACAGAAGCCCTAG 9600
 T  P  T  T  Q  L  S  M  N  M  G  L  A  V  P  L  W  L  A  T  V  L  I  G  L  Q  K  K  P  T  E  A  L

CCCACTTATTACCAGAAGGTACCCCAGCAGCACTCATTCCCATATTAATTATCATTGAAACTATTAGTCTTTTTATCCGACCTATCGCCCTAGGAGTCCG 9700
A  H  L  L  P  E  G  T  P  A  A  L  I  P  M  L  I  I  I  E  T  I  S  L  F  I  R  P  I  A  L  G  V  R

ACTAACCGCTAATTTAACAGCTGGTCACTTACTTATACAACTAGTTTCTATAACAACCTTTGTAATAATTCCTGTCATTTCAATTTCAATTATTACCTCA 9800
  L  T  A  N  L  T  A  G  H  L  L  M  Q  L  V  S  M  T  T  F  V  M  I  P  V  I  S  I  S  I  I  T  S

CTACTTCTTCTATTACTAACAATTCTGGAGTTAGCTGTTGCTGTAATCCAGGCATATGTATTTATTCTACTTTTAACTCTTTATCTGCAAGAAAACGTTT 9900
 L  L  L  L  L  L  T  I  L  E  L  A  V  A  V  I  Q  A  Y  V  F  I  L  L  L  T  L  Y  L  Q  E  N  V
```

(Figure 3-2 continued)

```
COIII →
ATGTCCCACCAAGCTCATGCATACCACATGGTAGACCCAAGCCCCTGACCTCTAACCGGTGCTGGCGCCGCATTATTAATAACCTCTGGCCTAGCCATAT 10000
|M  S  H  Q  A  H  A  Y  H  M  V  D  P  S  P  W  P  L  T  G  A  G  A  A  L  L  M  T  S  G  L  A  M
Y  V  P  P  S  S  C  M  P  H  G  *|

GATTTCATAAAAACTCCTGTATCTTAATAACACTTGGTCTAATCCTTATACTTCTTACAATATATCAATGATGACGAGACATTGTTCGAGAAGGCACCTT 10100
W  F  H  K  N  S  C  I  L  M  T  L  G  L  I  L  M  L  L  T  M  Y  Q  W  W  R  D  I  V  R  E  G  T  F

CCTTGGCCATCACACTTCACCAGTCCAACAAGGCCTTCGCTACGGAATAATCCTATTTATTATTTCAGAAGTTTGCTTTTTCGCAGGTTTCTTCTGAGCT 10200
  L  G  H  H  T  S  P  V  Q  Q  G  L  R  Y  G  M  I  L  F  I  I  S  E  V  C  F  F  A  G  F  F  W  A

TTCTATCATGCCAGTCTTGCACCAACCCCAGAACTTGGCTTAACATGACCCCCAACAGGAATTAACCCTCTAAACCCATTTGAAGTTCCACTATTGAATA 10300
 F  Y  H  A  S  L  A  P  T  P  E  L  G  L  T  W  P  P  T  G  I  N  P  L  N  P  F  E  V  P  L  L  N

CAGCTGTTTTACTTGCCTCAGGAGTTTCAGTAACTTGGGCCCATCACAGCATTACTGAAAAAAATCGAACAGAAACAACCCAAGCCCTAACTTTAACAGT 10400
T  A  V  L  L  A  S  G  V  S  V  T  W  A  H  H  S  I  T  E  K  N  R  T  E  T  T  Q  A  L  T  L  T  V

TTTACTAGGACTTTATTTTACTGCTCTGCAAATTATAGAATACTATGAAACCCCCTTTACAATAGCAGATGGCGTATACGGTTCAACATTCTTTGTCGCC 10500
  L  L  G  L  Y  F  T  A  L  Q  I  M  E  Y  Y  E  T  P  F  T  M  A  D  G  V  Y  G  S  T  F  F  V  A

ACAGGCTTTCACGGACTACATGTTATTATTGGCTCCCTATTCCTACTTACATGCTTACTACGACACTTACAATATCACTTCACCTCTAAACACCACTTCG 10600
 T  G  F  H  G  L  H  V  I  I  G  S  L  F  L  L  T  C  L  L  R  H  L  Q  Y  H  F  T  S  K  H  H  F

                                                                                                tRNA-Gly  →
GCTTCGAAGCCGCCGCTTGATACTGACACTTTGTAGACGTTGTGTGATTATTCCTATATATTTCAATCTACTGATGAGGCTCTTAACTCAGCCTGCCTTT 10700
G  F  E  A  A  A  W  Y  W  H  F  V  D  V  V  W  L  F  L  Y  I  S  I  Y  W  W  G  S  *|       |

                                                                ND3 →
TTAATACATTTAATATAGTTGGGTTCCAACCAACCAAACCTGGTATAAATCCAAGAAAAGGCACATGAACTCCTTTATAGTTATAATTATACTAACTCTA 10800
                                                                  ||M  N  S  F  M  V  M  I  M  L  T  L

ACCCTCTCATCTATTATAGCTCTTCTAGCATTTTGATTACCGATTATGAAACCAGACAGTGAAAAACTCTCTCCATACGAATGCGGATTCGACCCACAAG 10900
 T  L  S  S  I  M  A  L  L  A  F  W  L  P  I  M  K  P  D  S  E  K  L  S  P  Y  E  C  G  F  D  P  Q
```

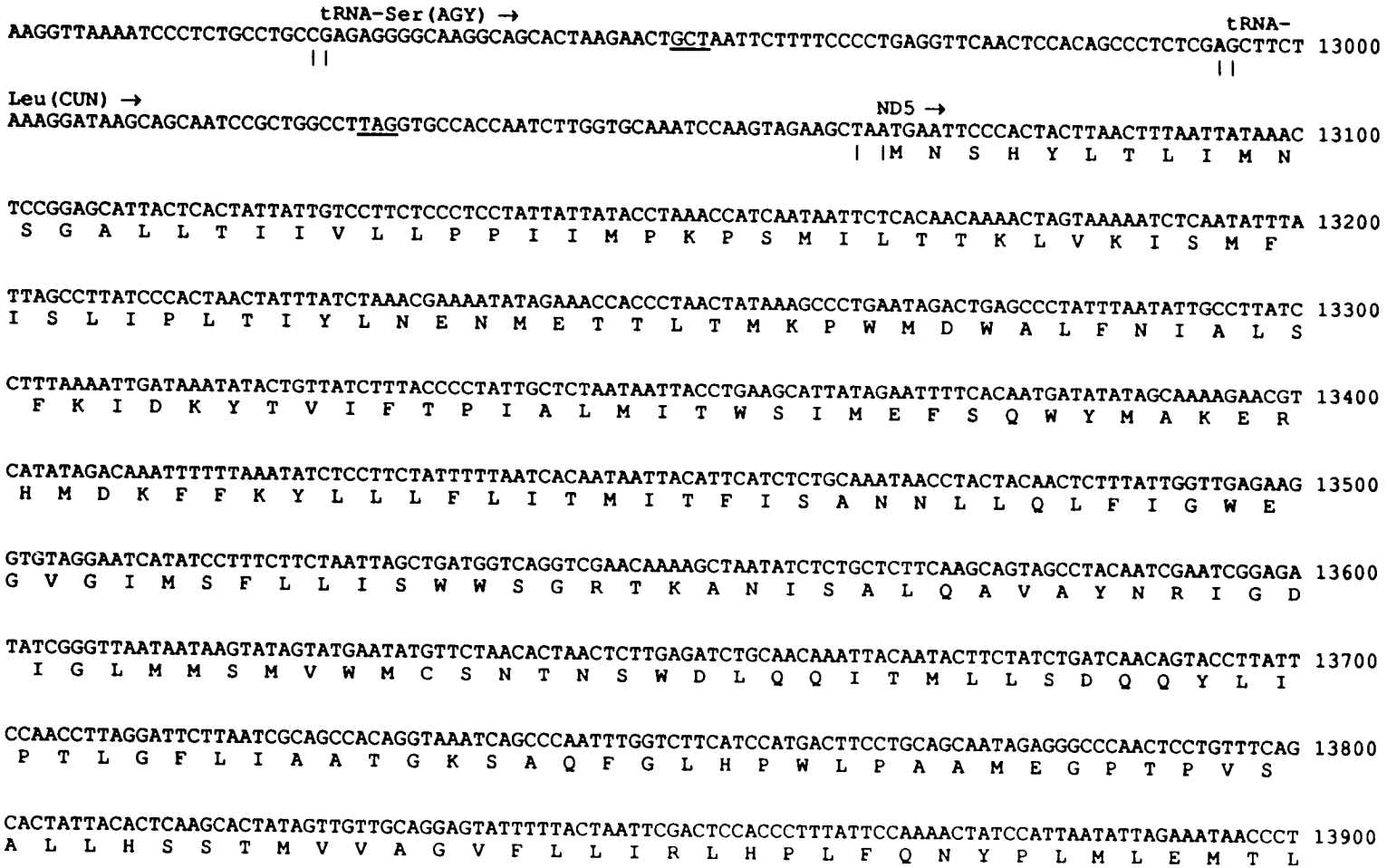
(Figure 3-2 continued)

```
GATCAGCCCGTCTACCCTTCTCTCTTCGATTCTTCTTAGTAGCAATCCTATTCTTATTGTTCGACTTAGAAATCGCCCTCCTTCTTCCATCCCCATGAGC 11000
G  S  A  R  L  P  F  S  L  R  F  F  L  V  A  I  L  F  L  L  F  D  L  E  I  A  L  L  L  P  S  P  W  A

AACTAATATTTCCAACCCAGAGTTCACCCTTCTCTGAGCTTCTTTATTTGTTTTACTTCTTACACTAGGACTAATCTATGAATGACTACAAGGAGGACTT 11100
  T  N  I  S  N  P  E  F  T  L  L  W  A  S  L  F  V  L  L  L  T  L  G  L  I  Y  E  W  L  Q  G  G  L

                  tRNA-Arg →                                                                    ND4L →
GACTGAGCAGAGTAATTTATTGGGGTTTAGTCTAATTAAGACAATTGATTTCGGCTCAATTAATCCTGAACTTTCAGGAACACCTACTCTCACATGCCTA 11200
 D  W  A  E  *  |     |                                                                  |        |M  P

CGACATTAATTTTTACCTCTTTTTTCCTGGCCTTATTAGGTCTCTCCCTGCAACGAAAACACCTTCTTTCACTCCTACTAACCTTAGAAAGTATGGCCCT 11300
T  T  L  I  F  T  S  F  F  L  A  L  L  G  L  S  L  Q  R  K  H  L  L  S  L  L  L  T  L  E  S  M  A  L

AGCATTATATGTTTCTACCGCACTATGAGCCTTAAACAACACATCCCTCCCAATTATAGCAGCCCCACTTATCATCTTAACCTTCTCAGCCTGTGAAGCT 11400
  A  L  Y  V  S  T  A  L  W  A  L  N  N  T  S  L  P  I  M  A  A  P  L  I  I  L  T  F  S  A  C  E  A

                                                                                  ND4 →
GGTATGGGTTTATCTCTAATAATTGCAACAGCTCGCACTCATAATACTGACCAACTAAAAGCACTAAACCTACTAAAATGTTAAAACTCATCATCCCTTC 11500
 G  M  G  L  S  L  M  I  A  T  A  R  T  H  N  T  D  Q  L  K  A  L  N  L  L  K|M  L  K  L  I  I  P  S
                                                                                C  *|

AATTATACTAATTCCAATAACCTTTTTAATTAACAAAAAAAGCTTACTATGAACTGCTACAACTTTCTTCAGCTTTTTAATCGCAGCTCTATCAACACTT 11600
  I  M  L  I  P  M  T  F  L  I  N  K  K  S  L  L  W  T  A  T  T  F  F  S  F  L  I  A  A  L  S  T  L

ACATTAAATATAGATGTAGCTGAACATAATTCAACCAATCCCCTTCTAAGCATTGACCAATTTTCATGCCCATTAATTATGCTATCTTGTTGACTTCTTC 11700
T  L  N  M  D  V  A  E  H  N  S  T  N  P  L  L  S  I  D  Q  F  S  C  P  L  I  M  L  S  C  W  L  L

CCCTAACTATCATAGGCAGTCAAGCACATATAAAAACTGAACCAATTACACGACAAAAGACAATAATTTCTCTACTTATTCTTCTCCAAGTCCTTCTATG 11800
P  L  T  I  M  G  S  Q  A  H  M  K  T  E  P  I  T  R  Q  K  T  M  I  S  L  L  I  L  L  Q  V  L  L  C

TATTACCTTCGGAGCCTCCAACCTACTTATATTCTATATCGCTTTCGAAACTACTTTAATCCCCACTCTTCTAATTATCACTCGTTGAGGTAACCAAAAG 11900
  I  T  F  G  A  S  N  L  L  M  F  Y  I  A  F  E  T  T  L  I  P  T  L  L  I  I  T  R  W  G  N  Q  K
```

(Figure 3-2 continued)

```
GAGCGACTCACAGCAGGCCTATATTTCCTATTCTACACTCTATCCGCCTCTCTCCCCCTCCTCCTTGCCCTTATCATAATTCAAACTCATTTAAACTCCT 12000
 E  R  L  T  A  G  L  Y  F  L  F  Y  T  L  S  A  S  L  P  L  L  L  A  L  I  M  I  Q  T  H  L  N  S

TATCAATCTATATTATTCCTCTATCTAATCTCACCCTATTATTAAACACACCTTGATCTGAAACCTTATGATGAATCGCCTGTTTCCTGGCCTTTTTAAT 12100
L  S  I  Y  Y  I  I  P  L  S  N  L  T  L  L  L  N  T  P  W  S  E  T  L  W  W  I  A  C  F  L  A  F  L  I

CAAAATACCCCTATATATCTTTCACTTATGATTACCAAAAGCTCACGTAGAGGCTCCCATCGCAGGCTCTATAATTCTAGCTGCAATTCTATTAAAACTA 12200
  K  M  P  L  Y  I  F  H  L  W  L  P  K  A  H  V  E  A  P  I  A  G  S  M  I  L  A  A  I  L  L  K  L

GGAGGTTACGGTATAATTCGTATATCATCTTTATTTATTCCACTAACTAAAGATCTGGCTGTCCCATTCATAATTATCGCCATATGAGGGATAATCGTAA 12300
 G  G  Y  G  M  I  R  M  S  S  L  F  I  P  L  T  K  D  L  A  V  P  F  M  I  I  A  M  W  G  M  I  V

CTAGTTCAATTTGTCTACGACAAACAGATCTAAAATCTATAATCGCTTACTCGTCTGTCAGCCATATAGGCCTAGTCGTAGCCGGCATTTTCACAATAAC 12400
T  S  S  I  C  L  R  Q  T  D  L  K  S  M  I  A  Y  S  S  V  S  H  M  G  L  V  V  A  G  I  F  T  M  T

TCCATGAGCATGATCTGGGGCTCTTGCAATAATAATTGCCCATGGATTAGTATCATCAGGTCTATTGTGTCTCGCTAATATTACATATGAACGCACTCAT 12500
  P  W  A  W  S  G  A  L  A  M  M  I  A  H  G  L  V  S  S  G  L  L  C  L  A  N  I  T  Y  E  R  T  H

ACACGTTCTATCTTCATAAACCGAGGTTTAAAAACTTTATTCCCTCTAATATCATTCTGATGACTTATAATAACTTTCGCCAATATAGCACTACCACCAT 12600
 T  R  S  I  F  M  N  R  G  L  K  T  L  F  P  L  M  S  F  W  W  L  M  M  T  F  A  N  M  A  L  P  P

TCCCAAACTTCATGGCAGAAATTTTAATCATTACCTCCTTATTTAACTGATCAAACTGAACCATCTTACTACTAGGGCTAAGCATAACCTTAACTGCCCT 12700
F  P  N  F  M  A  E  I  L  I  I  T  S  L  F  N  W  S  N  W  T  I  L  L  L  G  L  S  M  T  L  T  A  L

TTTCTCATTAAATATACTTATTATAACTCAACATGAACACCCAAATAAACATGCACCAGTTAACCCAAGTACCACCCGTGAACACCTACTTATACTTATA 12800
  F  S  L  N  M  L  I  M  T  Q  H  E  H  P  N  K  H  A  P  V  N  P  S  T  T  R  E  H  L  L  M  L  M

                                                                    tRNA-His →
CACATAGCCCCTATTATCCTTCTCATTGCTAACCCAAGCGCTATTATAATTAGAAGCACGCATAGTTTATACAAAACATTAGATTGTGAGTCTAATAAAG 12900
H  M  A  P  I  I  L  L  I  A  N  P  S  A  I  M  I  *| |
```

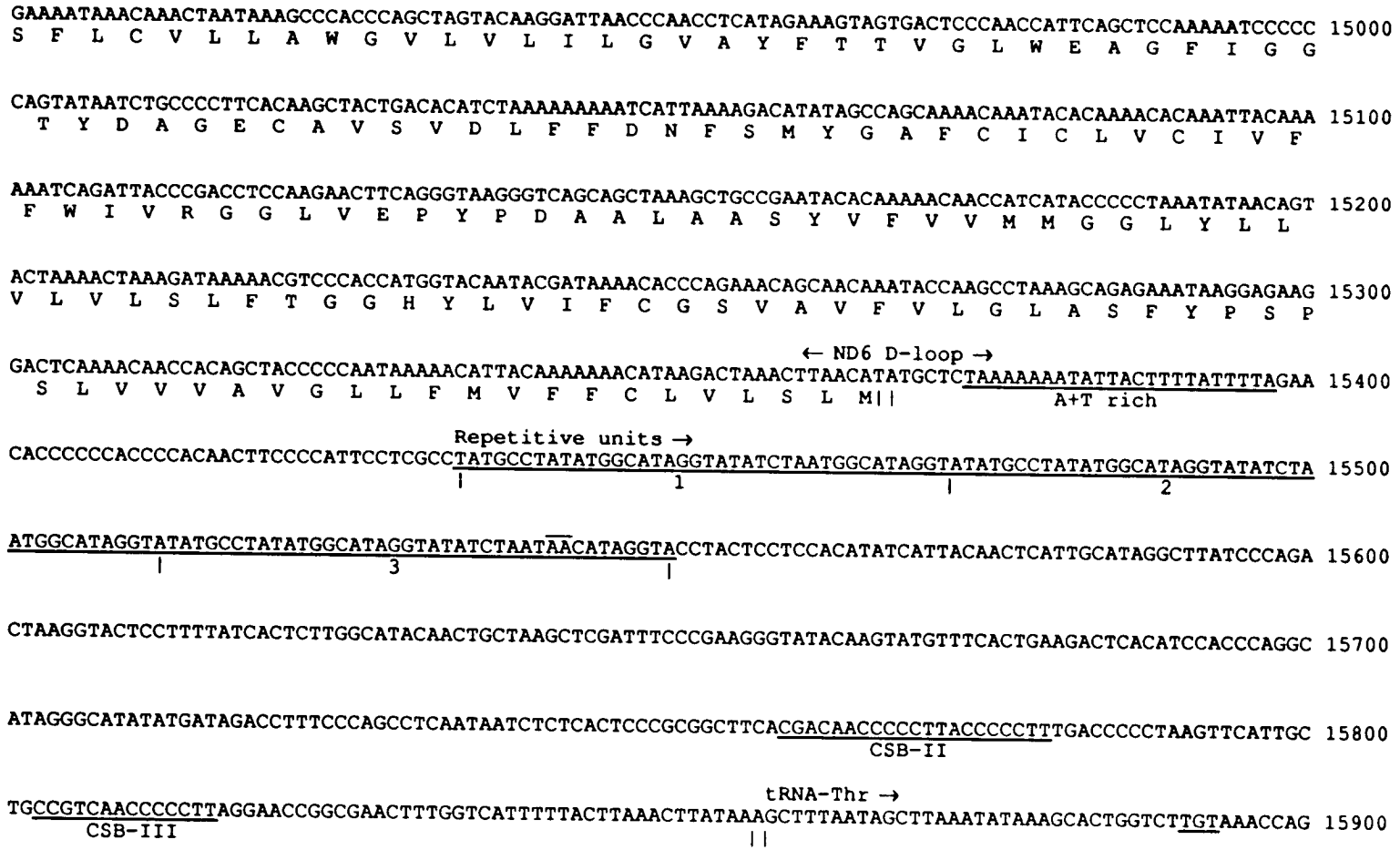
(Figure 3-2 continued)



(Figure 3-2 continued)

```
ATGTCTAGGAGCAATAACCACCATTTGTGCTGCCCTATGTGCAACAACACAAAATGATATCAAAAAAATTATTGCCTTTTCTACATCAAGTCAACTAGGC 14000
  C  L  G  A  M  T  T  I  C  A  A  L  C  A  T  T  Q  N  D  I  K  K  I  I  A  F  S  T  S  S  Q  L  G

TTAATAATAGTCGCAGTTGGTCTTAACCACCCTCACATTGCCTTTCTCCACATGTGTACACATGCCTTTTTTAAAGCTATACTTTTCTTATGCTCAGGAA 14100
 L  M  M  V  A  V  G  L  N  H  P  H  I  A  F  L  H  M  C  T  H  A  F  F  K  A  M  L  F  L  C  S  G

GTATTATTCATAATATAAATAATGAACAAGATATTCGAAAATTTAGCTGTTTAAATAACAACTTACCTCTAACAACAACCTGTATAACAATTGGGTCCGC 14200
S  I  I  H  N  M  N  N  E  Q  D  I  R  K  F  S  C  L  N  N  N  L  P  L  T  T  T  C  M  T  I  G  S  A

AGCACTAATAGGCTTACCATTTCTAGCTGGTTTCTTCACTAAAGACTTAATTCTAGAAGCCCTAAATACTTCCTATACTAATGCCTGAGCCCTAATAGTT 14300
  A  L  M  G  L  P  F  L  A  G  F  F  T  K  D  L  I  L  E  A  L  N  T  S  Y  T  N  A  W  A  L  M  V

ACTCTTATAGCCGTTACATTAACAACTGCCTACAGTTCACGCCTTATTATTATATCAGCCTCTGGTACACCACGATACTTACCCCTAACCCCAACACACG 14400
  T  L  M  A  V  T  L  T  T  A  Y  S  S  R  L  I  I  M  S  A  S  G  T  P  R  Y  L  P  L  T  P  T  H

AAAATAATTTTATCAAGAACCCATTAAAACGTTTAGCCTGGGGCAGCCTAATTTCAGGACTAATCCTTACAAGTACCCTCCCACCAATAAAACCTCAAAT 14500
E  N  N  F  I  K  N  P  L  K  R  L  A  W  G  S  L  I  S  G  L  I  L  T  S  T  L  P  P  M  K  P  Q  I

TTTTACAATACCAACCTATATTAAAACTATTGCTCTAATAATATTTATCATTAGCCTAATTATTTCTATAGAACTAACCAATAAAAAAATTAACCAAACT 14600
  F  T  M  P  T  Y  I  K  T  I  A  L  M  M  F  I  I  S  L  I  I  S  M  E  L  T  N  K  K  I  N  Q  T

ACATTCTCCTTCTTTACTCAACTAGCATTCTACCCCCATATTATCCATCGTTTAACATCCCACCTATCTTTAATCTGAAGTCAAAAATTAATAACACAAG 14700
 T  F  S  F  F  T  Q  L  A  F  Y  P  H  I  I  H  R  L  T  S  H  L  S  L  I  W  S  Q  K  L  M  T  Q

TAATAGACGTATCATGACTTGAAAAAATCGGACCAAAAGGTTTAGCTAATCACCAACTTAAACCCTCCACTACACTAACAGAAGCACATCACCTAAATTC 14800
V  M  D  V  S  W  L  E  K  I  G  P  K  G  L  A  N  H  Q  L  K  P  S  T  T  L  T  E  A  H  H  L  N  S

TGCCACCCTACCTTTAATAGCCTTTGCCCTAACCTTAATTACACTAAGTCTCACAGCTCGTAGAGCCCCACGATTTACCCCTCGAACAACTACCAAAACA 14900
  A  T  L  P  L  M  A  F  A  L  T  L  I  T  L  S  L  T  A  R  *|  G  R  N  V  G  R  V  V  V  L  L
                                                                |*  E  A  R  L  A
```

(Figure 3-2 continued)



(Figure 3-2 continued)

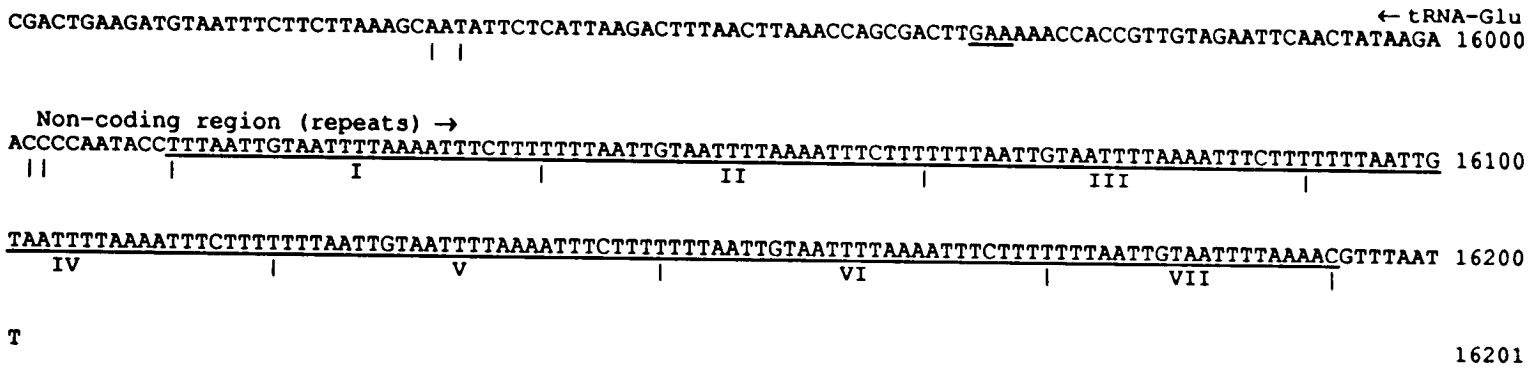


Table 3-2. Comparisons of lengths in base pairs of animal mitochondrial genes.

	Human	Mouse	Chicken	Frog	Loach	Sea Lamprey	Sea Urchin	*Drosophila*
D-loop	1043	879	1227	2134	896	491	121	1077
12S rRNA	954	975	819	951	989	900	976	867
16S rRNA	1559	1582	1621	1621	1680	1621	1530	1326
Cytb	1141	1144	1140	1140	1141	1191	1157	1137
ND1	956	957	975	970	975	966	969	975
ND2	1042	1036	1038	1039	1047	1044	1059	1025
ND3	346	345	348	342	351	351	351	354
ND4	1378	1378	1377	1384	1383	1377	1380	1339
ND4L	297	294	294	297	297	291	294	290
ND5	1811	1824	1818	1815	1837	1797	1914	1720
ND6	528	519	519	513	522	519	495	525
COI	1541	1545	1548	1549	1551	1554	1554	1536
COII	684	684	684	688	691	690	690	685
COIII	784	784	783	781	768	786	783	789
ATP6	679	681	681	679	684	714	690	674
ATP8	207	204	165	168	168	168	168	162
Total	16569	16295	16775	17553	16558	16201	15650	16019

respectively. Both repetitive sequences fold into highly stable secondary structures. In the tRNA cluster in which the second-strand replication is normally initiated, there is no non-coding segment in the lamprey genome. Thus, it is possible that the repeats in the second of the major non-coding regions may function as the second strand origin (O_2), because of the homology of the structure. In fact, the replication of the second strand initiates near the control region in *Drosophila* mtDNA (Clary et al., 1985). As another possibility, one of the tRNA genes between ND2 and COI could play a dual role.

Several intergenic sequences in other regions are found. Most of them are less than five nucleotides in length. However there is an unusually long segment (23 bp) between genes for ND1 and tRNA-Ile. This sequence does not fold into a secondary structure, so it is unlikely to function as the replication origin of the second strand.

There are six cases of gene overlap in this genome. The largest overlap occurs between the 3' end of ATP 6 and COIII (35 nucleotides). It is interesting that no protein-coding genes which are immediately followed by tRNA genes encoded on the same strand overlap. The only exception is that two bases of the tRNA-Trp gene are used for the stop codon of ND2. Adjacent protein-coding genes always overlap if no tRNAs are present between them (See Table 3-1 and Figure 3-2). This observation strongly supports the idea that the tRNA genes present between peptide genes function as signals for the initiations or terminations of translations (Ojala et al., 1981).

Protein-Coding Genes

Thirteen peptide genes are identified by their sequence homology to other vertebrate mtDNAs. A translation using the mammalian mitochondrial code yields proteins with lengths similar to that of other vertebrates (Table 3-2). This suggests that the same code is used in the lamprey (Table 3-3). One exception is observed in Cyt b gene which has longer sequence than any other animal mtDNAs sequenced so far. By comparison with loach Cyt b sequence, it is realized that lamprey Cyt b is 4 and 12 amino acids longer at the 5' and 3'-end respectively. The sea lamprey mtDNA uses ATG as translational initiation codons for 12 protein-coding genes and GTG for COI, which is identical to the loach and chicken mt genomes (Table 3-4). Using GTG at the beginning of open reading frames in mitochondria is not unusual, since the rat and sea urchin mt genomes also use GTG as initiation codon for ND1 and ATP8 respectively. All sense codons, except GCG, are used but with different frequencies in the lamprey mt genome. GCG is used for Ala residue in other animal mtDNAs, but the sea lamprey mt genome does not have GCG codon (Table 3-3). The protein-coding genes in the lamprey mt genome end with three different termination codons. Six genes (Cyt b, ND4, ND5, ND6, ATP6, and COI) employ AGA as termination codons while another six genes (NDI, ND3, ND4L, ATP8, COII, and COIII) use TAA and ND2 uses TAG. Because of gene overlap, one nucleotide is read by two reading frames in the case of ATP8/ATP6, ATP6/COIII, ND4L/ND4, and ND5/ND6. One-nucleotide frameshifts are observed in the overlapping gene pairs. In

Table 3-3. Genetic code and codon usage from all protein-coding genes in the sea lamprey mitochondrial genome.

Code (AA)	# (fraction)	Code (AA)	#(fraction)	Code(AA)	# (fraction)	Code (AA)	#(fraction)
TTG (Leu)	12 (.02)	TCG (Ser)	3 (.01)	TAG (End)	2 (.15)	TGG (Trp)	10 (.09)
TTA (Leu)	199 (.33)	TCA (Ser)	88 (.33)	TAA (End)	5 (.38)	TGA (Trp)	99 (.91)
TTT (Phe)	137 (.58)	TCT (Ser)	73 (.28)	TAT (Tyr)	69 (.64)	TGT (Cys)	27 (.68)
TTC (Phe)	98 (.42)	TCC (Ser)	44 (.17)	TAC (Tyr)	39 (.36)	TGC (Cys)	13 (.32)
CTG (Leu)	15 (.02)	CCG (Pro)	4 (.02)	CAG (Gln)	6 (.06)	CGG (Arg)	2 (.03)
CTA (Leu)	194 (.32)	CCA (Pro)	113 (.56)	CAA (Gln)	92 (.92)	CGA (Arg)	37 (.50)
CTT (Leu)	137 (.22)	CCT (Pro)	53 (.26)	CAT (His)	48 (.45)	CGT (Arg)	21 (.28)
CTC (Leu)	53 (.09)	CCC (Pro)	32 (.16)	CAC (His)	58 (.55)	CGC (Arg)	8 (.11)
ATG (Met)	29 (.13)	ACG (Thr)	2 (.01)	AAG (Lys)	9 (.09)	AGG (End)	0 (.00)
ATA (Met)	188 (.87)	ACA (Thr)	125 (.41)	AAA (Lys)	90 (.91)	AGA (End)	6 (.46)
ATT (Ile)	232 (.69)	ACT (Thr)	106 (.35)	AAT (Asn)	78 (.54)	AGT (Ser)	33 (.13)
ATC (Ile)	102 (.31)	ACC (Thr)	73 (.24)	AAC (Asn)	67 (.46)	AGC (Ser)	22 (.08)
GTG (Val)	8 (.04)	GCG (Ala)	0 (.00)	GAG (Glu)	11 (.12)	GGG (Gly)	33 (.15)
GTA (Val)	73 (.38)	GCA (Ala)	99 (.33)	GAA (Glu)	82 (.88)	GGA (Gly)	85 (.40)
GTT (Val)	79 (.41)	GCT (Ala)	96 (.32)	GAT (Asp)	32 (.47)	GGT (Gly)	52 (.24)
GTC (Val)	32 (.17)	GCC (Ala)	102 (.34)	GAC (Asp)	35 (.55)	GGC (Gly)	43 (.20)

Table 3-4. Initiation and termination codons used in the animal mitochondrial protein-coding genes. The asterisk (*) indicates the incomplete termination codons.

	Cytb	ND1	ND2	ND3	ND4	ND5	ND6	ND4L	ATP6	ATP8	COI	COII	COIII
Human	ATG/*	ATA/*	ATT/*	ATA/*	ATG/*	ATA/TAA	ATG/AGG	ATG/TAA	ATG/TAA	ATG/TAG	ATG/AGA	ATG/TAG	ATG/*
Cow	ATG/AGA	ATG/*	ATA/*	ATA/*	ATG/*	ATA/TAA	ATG/TAA	ATG/TAA	ATG/*	ATG/TAA	ATG/TAA	ATG/TAA	ATG/*
Mouse	ATG/*	ATT/*	ATA/*	ATC/TAA	ATG/*	ATC/TAA	ATG/TAA	ATG/TAA	ATG/*	ATG/TAA	ATG/TAA	ATG/TAA	ATG/*
Rat	ATG/TAA	ATT/*	ATA/*	ATT/TAA	ATG/*	ATA/TAA	ATG/TAA	ATG/TAA	ATG/*	ATG/TAA	ATG/TAA	ATG/TAA	ATG/*
Chicken	ATG/TAA	ATG/TAA	ATG/TAG	ATG/TAA	ATG/*	ATG/TAA	ATG/TAA	ATG/TAA	ATG/*	ATG/TAA	ATG/TAA	ATG/TAA	ATG/*
Frog	ATG/TAG	ATG/*	ATG/TAG	ATG/TGA	ATG/*	ATG/TAA	ATG/AGA	ATG/*	ATG/TAA	ATG/TAA	GTG/AGG	ATG/TAA	ATG/*
Loach	ATG/*	ATG/TAA	ATG/TAG	ATG/TAG	ATG/TAG	ATG/TAA	ATG/TAA	ATG/TAA	ATG/TAA	ATG/TAA	ATG/*	ATG/*	ATG/*
Lamprey	ATG/AGA	ATG/TAA	ATG/TAG	ATG/TAA	ATG/AGA	ATG/AGA	ATG/AGA	ATG/TAA	ATG/AGA	ATG/TAA	GTG/TAA	ATG/*	ATG/TAA
Urchin	ATA/TAG	ATG/TAA	ATG/TAG	ATG/TAA	ATA/TGA	ATG/AGA	ATG/AGA	ATC/TAA	ATA/TAA	GTG/TAA	ATG/TAA	ATG/TAA	ATG/TAA
Drosophila	ATG/TAA	ATA/TAA	ATT/TAA	ATT/TAA	ATG/*	ATT/*	ATT/TAA	ATG/TAA	ATG/TAA	ATT/TAA	ATA/TAA	ATG/*	ATG/TAA
Nematode	TTG/TAA	ATA/*	TTG/TAA	ATT/TAA	TTG/TAA	ATT/TAA	ATA/TAA	ATT/TAG	-	ATT/TAA	ATT/TAG	ATT/TAA	ATA/TAA

human mtDNA, only one mRNA is found in ATP8/ATP6 and ND4L/ND4, respectively (Anderson et al., 1981). It is possible that the lamprey genome has the same transcriptional process for ATP8/ATP6 and ND4L/ND4 genes. Although the structure of transcripts for the other overlapping genes is not known, they may also have one mRNA for each pair.

Base Composition -Bias and Skew

The base composition of mitochondrial genomes varies across animal taxa. Moreover, the nucleotide composition between the two strands is largely heterogeneous (Perna and Kocher, 1994). The wobble positions of four-fold degenerate sites, which may be relatively free of selective constraints on base substitution, best reflect the mutational spectra of the genome and therefore may be useful for understanding the underlying pattern of evolution of the mitochondrial protein-coding genes. In fact, the base composition of four-fold degenerate sites varies among deuterostomes, and usually differs from the base composition of the whole genomes (Table 3-5).

The frequency of four nucleotides in the whole lamprey genome appears to be homogeneous among deuterostome animals (Table 3-5). However, the base composition at the third positions of four-fold degenerate codon families is different from that of other vertebrate mtDNAs. The most exceptional base composition at the wobble positions is the frequency of Ts (32.7%), which is the highest among

Table 3-5. Base compositions of the wobble positions of 4-fold degenerate codons of all protein-coding genes located on the first strand, and of whole sequences of the same strand among deuterostome mtDNAs. The base composition of 4-fold degenerate sites is heterogeneous and does not always reflect the total base composition.

	Four-fold sites (%)				Whole genome (%)			
	A	C	G	T	A	C	G	T
Sea urchin	38.5	23.9	12.6	25.0	28.7	22.7	18.4	30.2
Lamprey	43.2	20.5	3.5	32.7	32.2	23.8	13.5	30.4
Frog	45.5	20.7	5.5	28.3	33.1	23.5	13.5	30.0
Mouse	54.8	21.3	4.6	20.0	34.5	24.4	12.3	28.7
Rat	52.5	28.7	3.3	15.5	34.1	26.2	12.5	27.2
Cow	49.7	27.8	5.3	17.1	33.4	27.2	13.5	27.7

deuterostomes whose sequences are available, suggesting that the mutational spectra of the genome is biased toward Ts. The higher T frequency is common in the protostome and earlier ancestral animal mtDNAs. The percentage T in a nematode (*A. suum*) is 51.9, whereas the human mtDNA contains only 25.6% of T (Anderson et al., 1981; Okimoto et al., 1992).

The base frequency at the wobble positions must be subject to selective constraints such as translational efficiency beyond the base composition of the genome. Because mtDNA generally uses a single tRNA for each codon family except for Leu and Ser, tRNA abundance for most tRNAs is not a factor for codon usage. However, different codons may have different affinity for the anticodon, ultimately resulting in differences in translational efficiency.

In addition to the intergenomic variation, intragenomic variations of base composition also exist in the lamprey mitochondrial genome. As presented in Table 3-6, the base composition of four-fold degenerate codons among protein genes of the lamprey mt genome appears to be heterogeneous. The range of difference in %GC is from 35.6% (Cyt b) to 19.9% (COI). The differences indicate that the individual genes may use different mutational matrices. Even though the base substitutions in the wobble positions do not result in the substitutions of amino acids, the biased base frequency may be the result of varying selective constraints from gene to gene.

The difference in the base distribution between two mitochondrial strands is called 'skew' (Perna and Kocher, 1994). As usually observed in the animal

mitochondrial DNAs, the GC- and AT-skews are an opposite pattern (Table 3-6). The positive GC-skew and negative AT-skew exist in all protein genes located on the sense strand. The reversed pattern is found in ND6 gene which is located on the second strand. The GC-skew is observed from -1.00 in ATP8 to -0.60 in ND2 and ATP6 genes. The AT-skew ranges from -0.11 in ATP8 gene to 0.25 in ND5 gene. The extreme skew in ATP8 is probably due to the sampling error caused by the small size of this gene. Although no clear relationship between the bias and skew across genes was found (Fig. 3-3), generally AT- and GC-skews are parallel in most genes except COII and ND5. It has been proposed that the extent of GC-skew might be associated with the length of time being single-stranded during the replication process, because the single strand is more easily subject to deamination (Thomas et al., unpublished ms). Since the location of the second strand replication origin is unknown in the lamprey genome, it is impossible to see a direct relationship between the patterns of skews or bias and the time of being single-stranded. Whether the second strand origin is assumed to be in one of tRNA genes near where it is found in other vertebrates, or if it is located in the second non-coding region, no clear underlying pattern explaining the correlation between the bias and time of being single-stranded is observed.

Transfer RNA Genes

Twenty-two tRNA genes can be identified in the sea lamprey mt genome through analyses of sequence similarity and potential secondary structure. The size

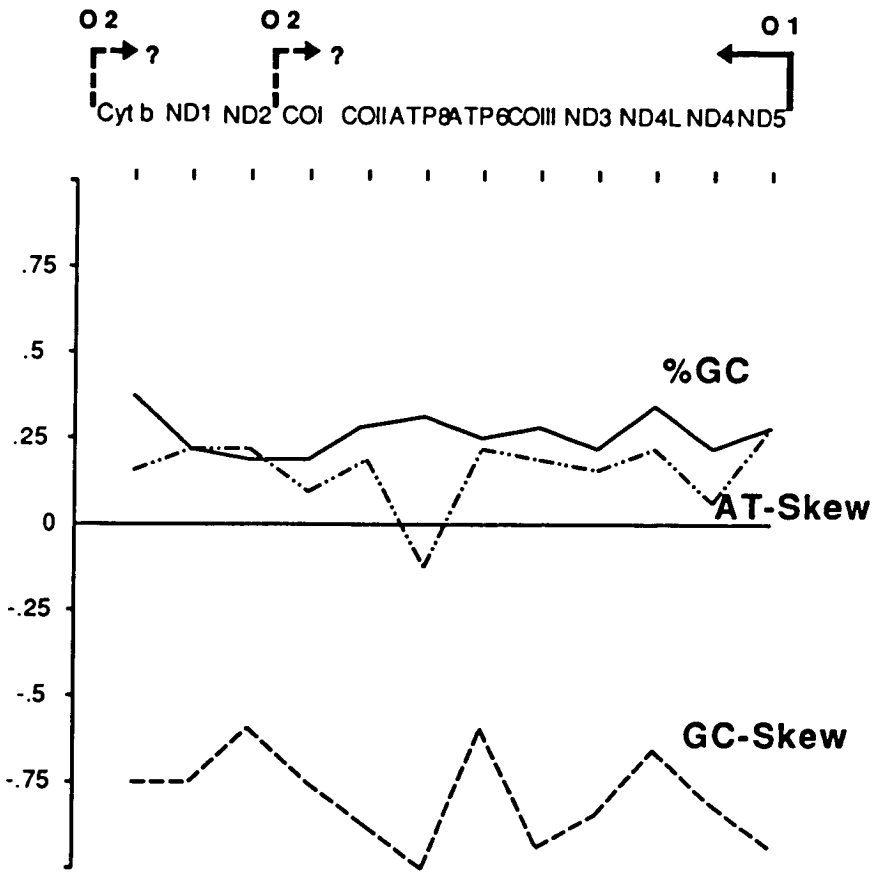


Figure 3-3. The patterns of %GC, GC-skew, and AT-skew in the lamprey mitochondrial genome. There is no clear underlying trend showing the relationship between the bias or skews and the locations of genes.

Table 3-6. Base composition of the third position of four-fold degenerate codons from all protein coding genes. The %GC, GC-, and AT-skews are heterogeneous across genes. The average is calculated without ND6.

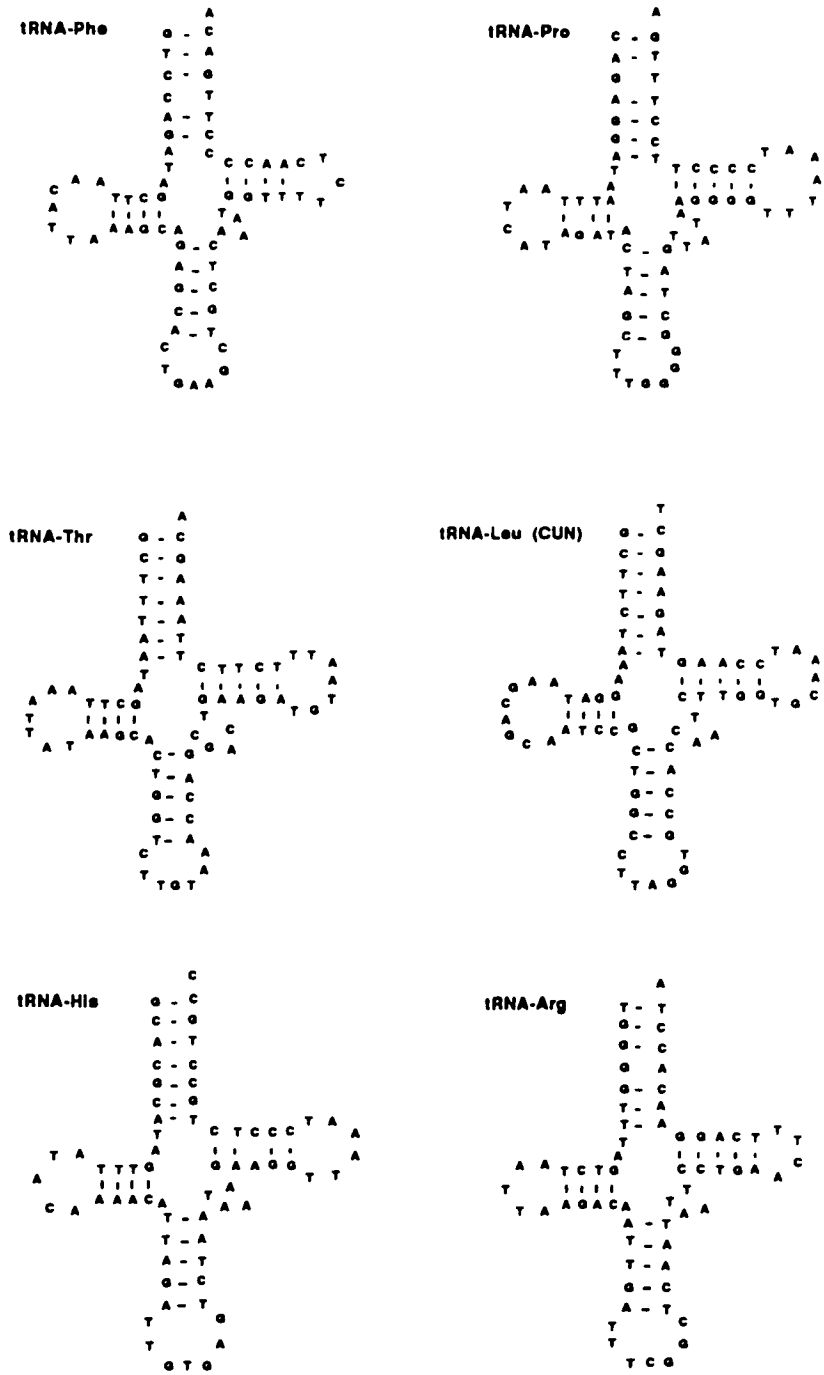
Gene	Cyt b	ND1	ND2	COI	COII	ATP8	ATP6
%GC	35.6	22.5	20.3	19.9	26.9	28.0	24.5
GC-skew	-0.74	-0.74	-0.60	-0.74	-0.86	-1.00	-0.60
AT-skew	0.13	0.24	0.21	0.10	0.19	-0.11	0.22

Gene	COIII	ND3	ND4L	ND4	ND5	Aver.	ND6
%GC	28.6	22.4	32.0	25.5	23.0	25.7	25.7
GC-skew	-0.90	-0.85	-0.67	-0.79	-0.91	-0.78	0.65
AT-Skew	0.18	0.16	0.21	0.09	0.25	0.15	-0.37

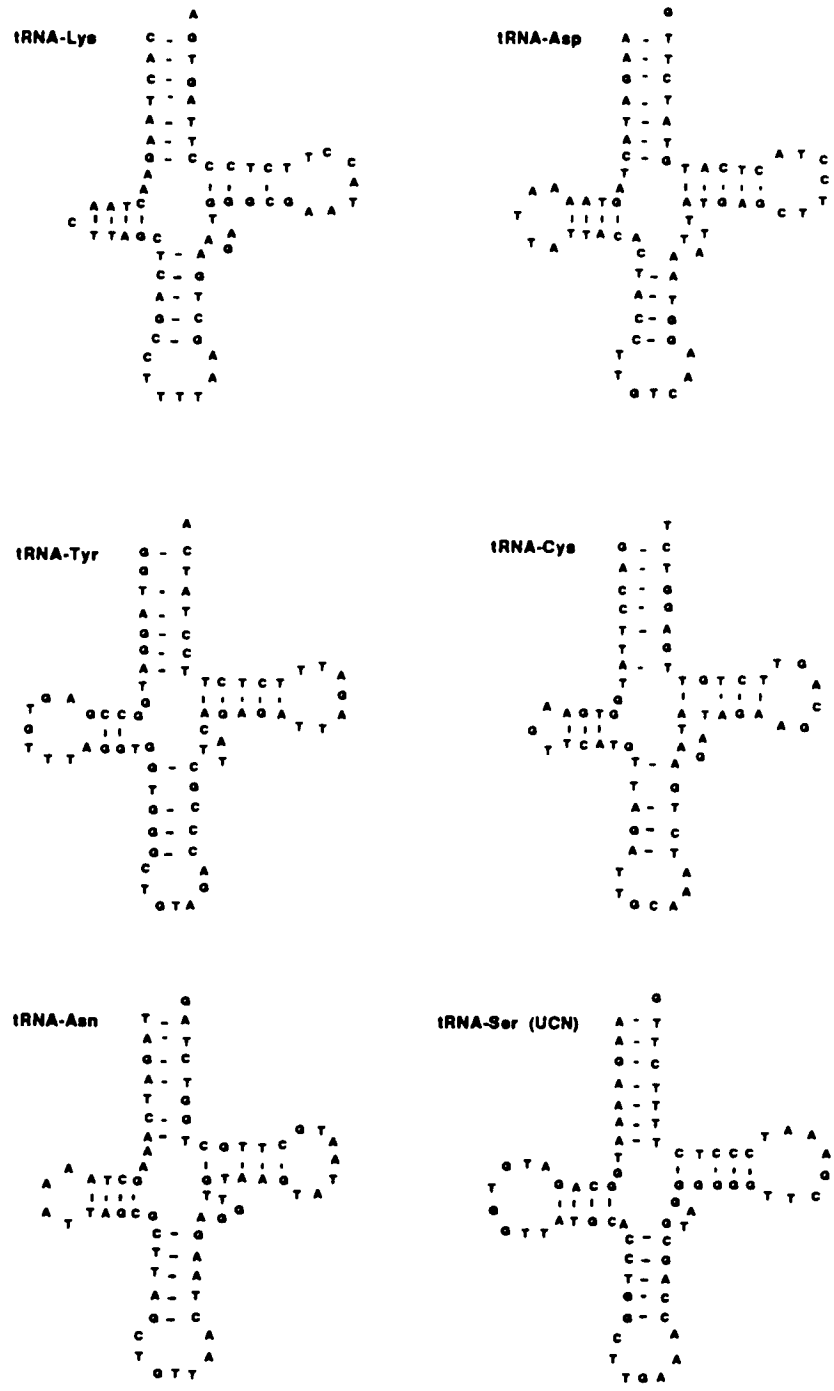
of lamprey tRNA genes is from 65 to 74 bases showing high variation in the dihydrouridine (DHU) and T-Ψ-C arms (Fig. 3-4). The DHU arm is the most variable in length. Most of them have 4 bp in the stems, but the loops vary in size. For instance, tRNA-Ser(AGY) has a very short DHU arm making it impossible to define the stem and loop due to the base mismatches. As another example, tRNA-Lys has only one base for the loop. T-Ψ-C and anticodon arms have 5 bp in the stem regions with some mismatches. However all anticodons are identical to other vertebrate mtDNAs.

The tRNA structures are very conserved with those of chicken and frog. By comparison with other animal mt tRNA genes, most of the base substitutions in the stem regions are compensatory changes, suggesting that a strong selective force maintaining the structure is operating and that the structure is very important for the functions of tRNAs. There are 24 U-G pairs found in the stem regions. These base matchings which are also found in tRNA genes from other animal mtDNAs and other ribosomal genes, and which are thought to be intermediate matchings before complete compensatory substitutions (Brown, 1985), may explain the initiation codon (GTG) for COI. Because the anticodon of tRNA-Met is CAU, the U can pair with G despite the possibly low affinity. Therefore the GTG also may be in transitional state before changing to a stable match to the anticodon.

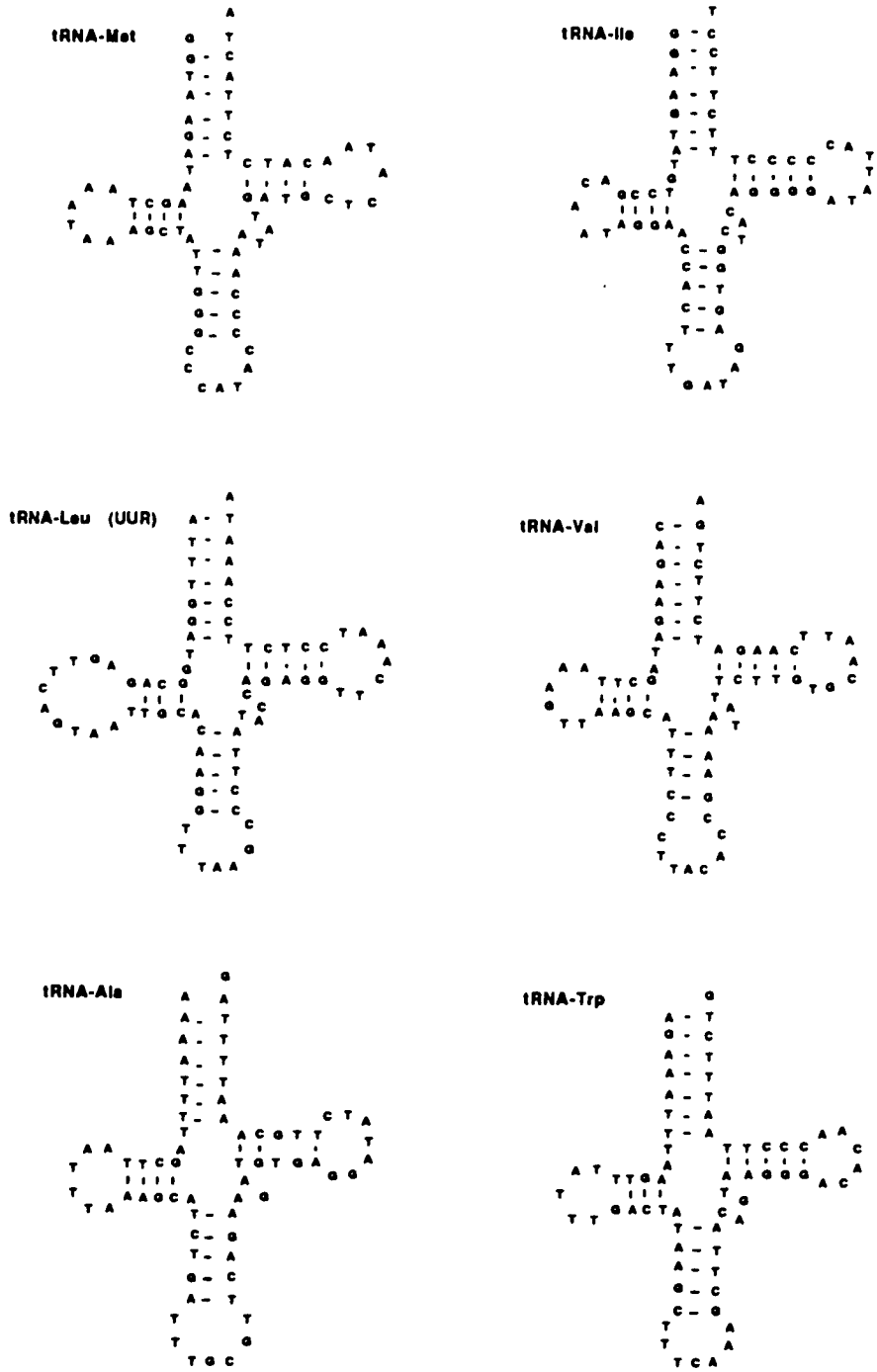
Figure 3-4. The secondary structures of the tRNA genes. Standard base matches are indicated by dashes. These tRNA genes show high sequence and structural variation in DHU stems.



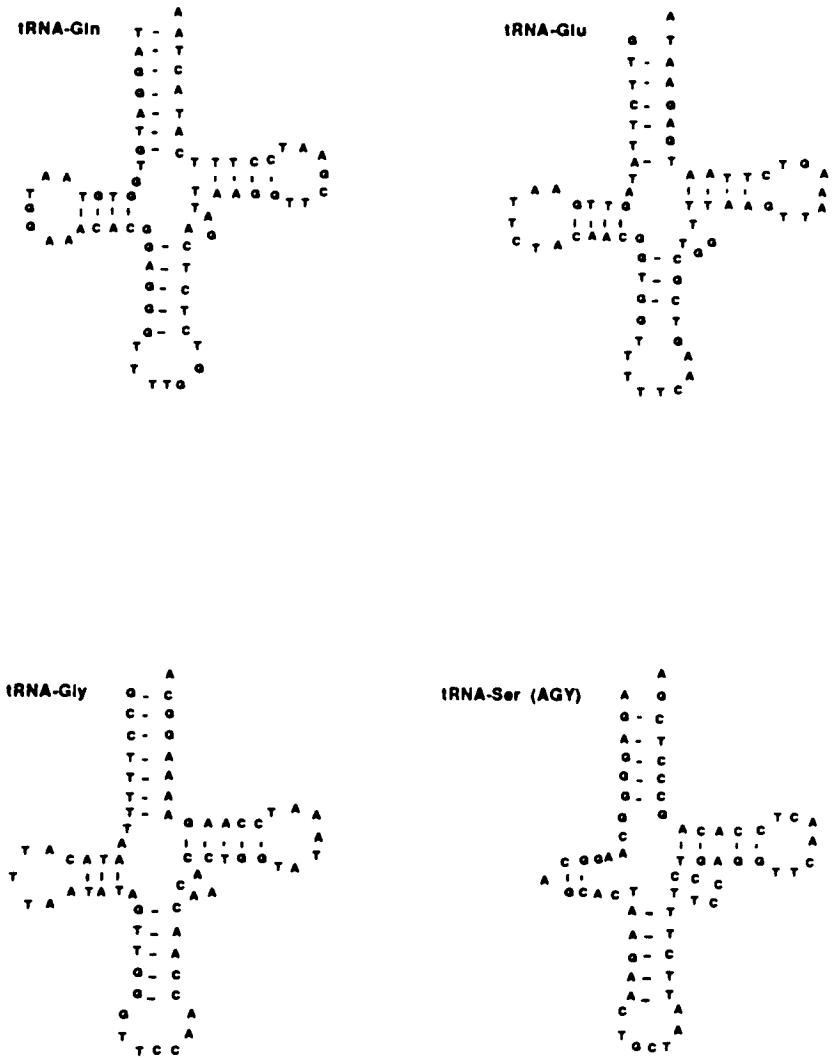
(Figure 3-4 continued)



(Figure 3-4 continued)



(Figure 3-4 continued)



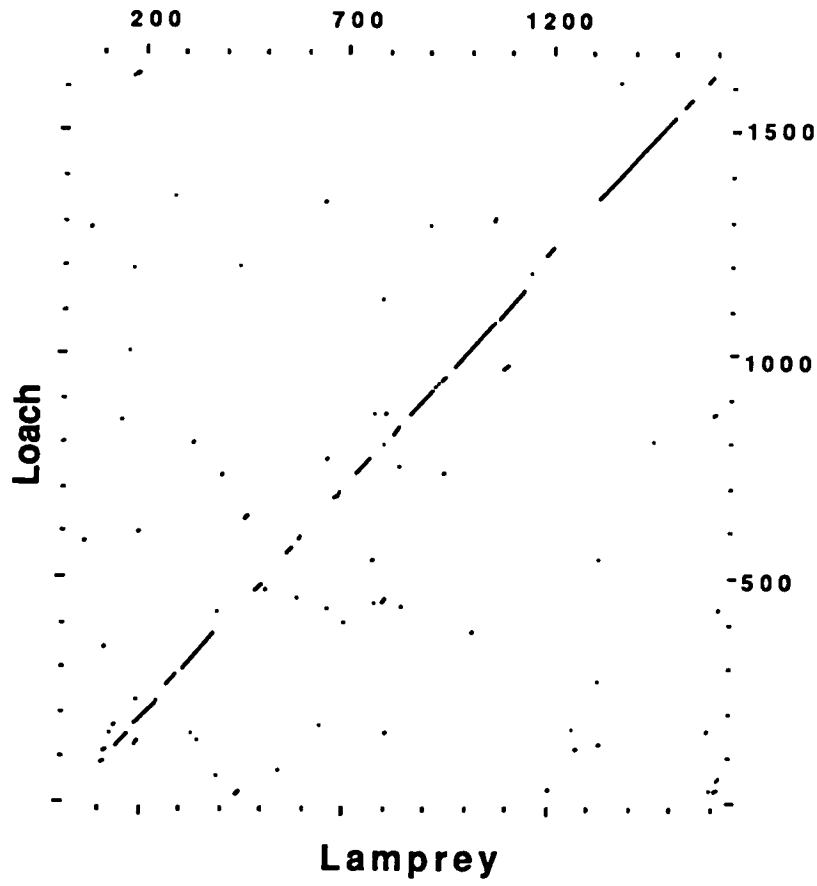
Ribosomal RNA Genes

Two ribosomal RNA genes are located between tRNA-Phe and tRNA-Leu (UUR) and separated by tRNA-Val as found in other vertebrate mt genomes. The sizes of 12S and 16S rRNA genes are 900 bp and 1621 bp in length respectively, showing little size variation. However, in comparison with loach ribosomal RNA genes, lamprey 12S and 16S are about 90 bp and 60 bp shorter respectively (Table 3-2). In 12S, the size variation between lamprey and loach is mainly from a long segment of indels in the 3'- ends while 16S has a long string of indels in the 5'-end. Several small size indels in the middle of the genes are observed in both genes.

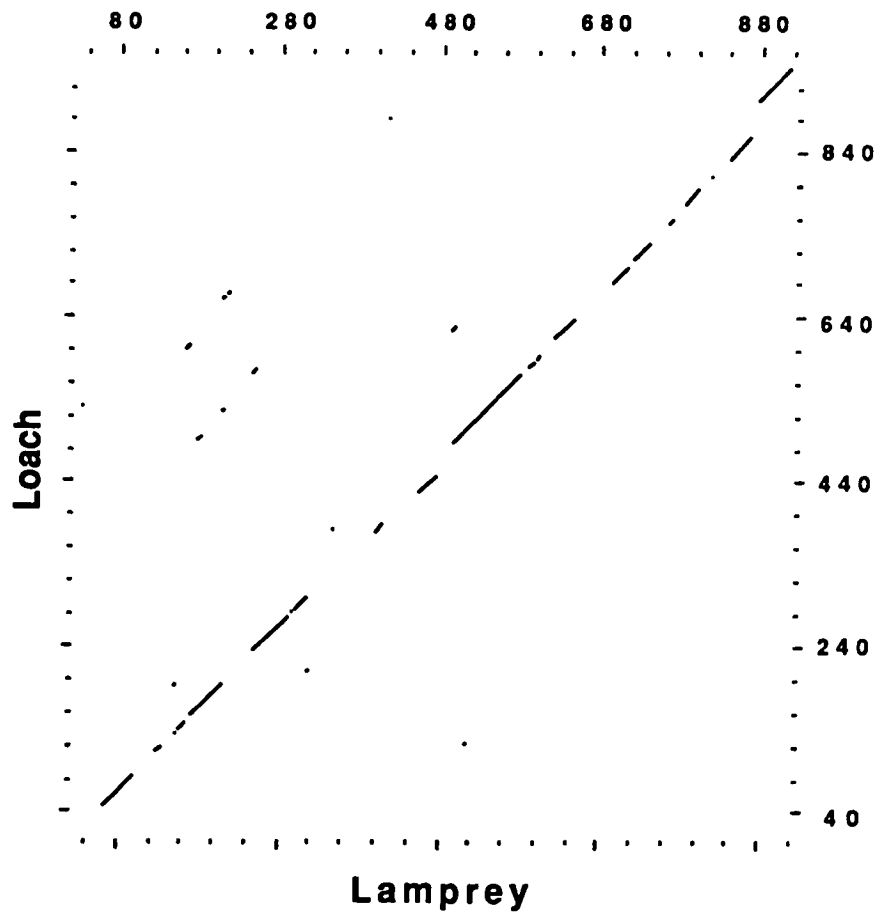
The 12S gene is more conserved than the 16S gene in sequence divergence. The sequence similarities between lamprey and loach ribosomal RNA genes are 64% and 53% in 12S and 16S respectively. Dotplots show that it is also common in the lamprey genome that rRNA genes have both conserved and variable internal regions (Fig. 3-5). The patterns of sequence divergence in the rRNA genes must be complementary to the structures of the genes, because the long single-stranded regions (loops) evolve slower than stem regions that accumulate compensatory base changes (Vawter et al., 1993). The secondary structure of 12S gene is relatively conserved with that of loach 12S gene. Since some regions are extremely conserved throughout animal taxa, I can expand the universality of the PCR primers to the population studies of lamprey species with the conserved sequence regions.

Figure 3-5. Comparisons of sequences of two rRNA genes. The horizontal sequence is from the lamprey mtDNA and the vertical sequence is from the loach mtDNA. The stringency used was 0.67. a) Dotplot comparison of 16S rRNA sequences. Several internal parts (especially, from 300 to 700 in the lamprey sequence) are highly variable. b) Dotplot comparison of 12S rRNA sequences. Many small conserved domains are observed.

A) 16S rRNA



B) 12S rRNA



Non-Coding Regions and Evolution of the Control Region

In the lamprey genome, the two major non-coding regions are separated by two tRNA genes, Thr and Glu. The tRNA-Thr is encoded on the H-strand, whereas the tRNA-Glu is on the L-strand.

The length of the first non-coding region is 491 bp including approximately 100 bp of tandem repeats. The first non-coding region contains a 28 bp string of an A+T rich (93%) region in 6 nucleotide downstream from the 5'-end. As marked in the Figure 3-4, three copies of a 39 bp sequence starting after 67 bp from 5'-end of the region are unique repeats. Besides the repeats, the first non-coding region consists of about 360 bp.

By the comparison of the control region sequences from mouse (Bibb et al., 1981), loach (Tzeng et al., 1992), and frog (Roe et al., 1985) with the 360 bp of the first non-coding region, I identified the conserved sequence block (CSB)-II and -III in the lamprey sequence. The lamprey CSB-II is exactly identical to the loach except one indel. Since Chang et al. (1986) reported that the CSB-II plays an important role for the initiation of heavy strand replication of mammalian mtDNAs, it is concluded that the first non-coding region is most likely the control region carrying most of regulatory sequences for replication, and the CSB-II is also important for lamprey mtDNA as well. If so, the control region of lamprey mtDNA is much shorter than that of any other vertebrate mtDNA.

Unexpectedly, none of the central conserved regions (usually marked A to F

in the mammalian mtDNAs, Kocher et al., 1991) is identifiable in the putative control region of lamprey mtDNA. Even though the CSB-D block is well conserved through bony fishes to mammals (Lee and Kocher, 1994), I could not identify any homology in the lamprey. The absence of the central conserved region in the lamprey suggests that the conserved central domain of mammals is a recently evolved domain.

In contrast to other vertebrate mtDNAs where two tRNAs (Pro and Phe) located 5' and 3'-ends of the control region supposedly play roles for initiation and termination of DNA synthesis (Jacobs et al., 1988), the lamprey mtDNA does not contain a tRNA gene next to the 5'-end of the putative control region. Instead there are two tRNA genes in the middle of two major non-coding regions. If the tRNAs (commonly Pro and Phe; Glu and Phe in birds) in both ends of the control region of other vertebrate mtDNAs are critical for the mt replicational metabolism, the lamprey mtDNA should have a substitute near the 5'-end. In this regard, I folded the repeats found in the putative control region into a secondary structure. The secondary structure is very stable (Fig. 3-6a), suggesting that the tandem repeat in the first non-coding region may have the same regulatory function as the tRNA does in other vertebrate mtDNAs or it is closely related with the termination-associated sequences.

Many studies have reported tandem repeats found in the control region, which are capable of forming highly stable secondary structures (Arnason et al., 1992; Wilkinson et al., 1991). Although the origin of repetitive sequences and the precise mechanisms giving rise to the repeats are unknown, it has been proposed that if the

tandem repeats contain large motifs, usually about 20 bp, and they have small motifs within them, the repeats might arise by DNA slippage during replication. After the slipped-DNA strand mispairings, deletion events are commonly observed in the eukaryotic genomes (Levinson et al., 1987). The larger motifs are suggested to arise by a interhelical mechanism such as unequal crossing over and genetic conversion (Hoelzel, 1993). The motif of this repeat is 39 bp, which is fairly long and may be within the range of long motifs. However, because there are two submotifs (12 bp) within the 39 bp unit, it probably arose by simple DNA slippage. Moreover, the repeats might have originated long time ago because the slippage or duplication must occur several times to generate the present motif. The motif of repeats is as follows:

5'-TATGCCTATATGGCATAGGTATATCTAATGGCATAGGTA-3'

Within the repetitive motif, two subunits of a repeat sequence are underlined. In the last unit of the repeats, there are two transitional substitutions (G → A), which are double underlined. In short, the internal sequence of the repeats suggest that it evolved in a stepwise fashion from a shorter fundamental repeat. The schematic diagram for the processes is as follows:

Let ATGGCATAGGTA be X.

X → (XX) → [(XX)(XX)] → [(XX)(XX)][(XX)(XX)] → (deletion) (XX)(XX)(XX)

The second non-coding region is 199 bp in length, and consists nearly entirely of seven copies of a 27 bp string. The tandem repeats do not have sequence variation. Unlike the repeats in the first non-coding region, these units show highly

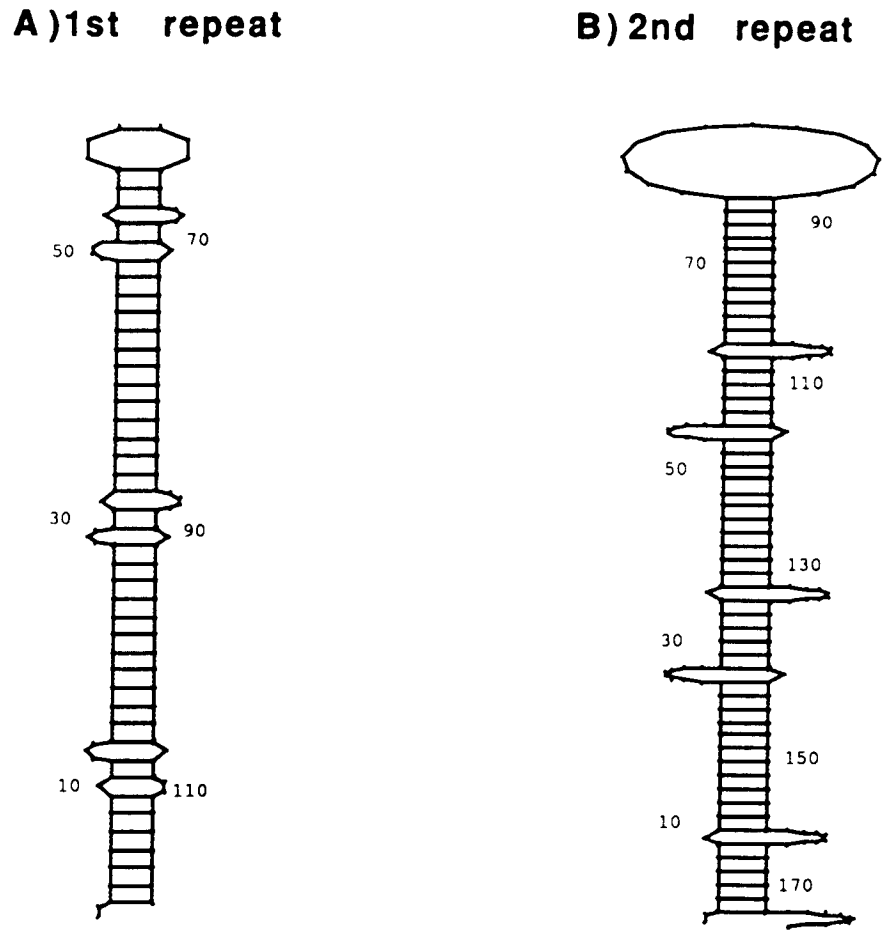


Figure 3-6. The secondary structures of tandem repeats in the two major non-coding regions. Both secondary structures are highly stable long stems. There are two transitional substitutions in the first repeats, no variation was found in the second repeats.

A+T rich base composition (93%) as follows.

5'-TTTAATTGTAATTTTAAAATTTCTTTT-3'

These tandem repeats also form highly stable secondary structures (Fig. 3-6b). Therefore both repeats in the first and second non-coding regions might have arisen by the same mechanistic process but the roles of those repeats may be different. As found in flounders (Lee and Kocher, 1994) and a marsupial (Janke et al., 1994), the repeats in the 3'-end of the control regions greatly expand the sequences of the control region. In this regard, the repetitive sequences in the second non-coding region may be responsible for expanding the sequence of the lamprey mitochondrial genome, while the first one may function as a subset of the regulatory sequence.

Evolution of Mitochondrial Gene Order

The lamprey shows several changes in gene order near the non-coding regions, relative to other vertebrates (Fig. 3-7). The first major non-coding region is located between ND6 and tRNA-Thr and is probably homologous to the control region of other vertebrates, since it contains sequences similar to CSB-II and CSB-III. The second non-coding region, located between tRNA-Glu and Cyt b, consists almost entirely of repeated sequence. Two tRNA genes (Pro and Phe) which are normally adjacent to both ends of the control region, are instead located next to each other between Cyt b and 12S, but tRNA-Phe is still next to 12S.

Two major mechanisms have been proposed for changes in mitochondrial gene

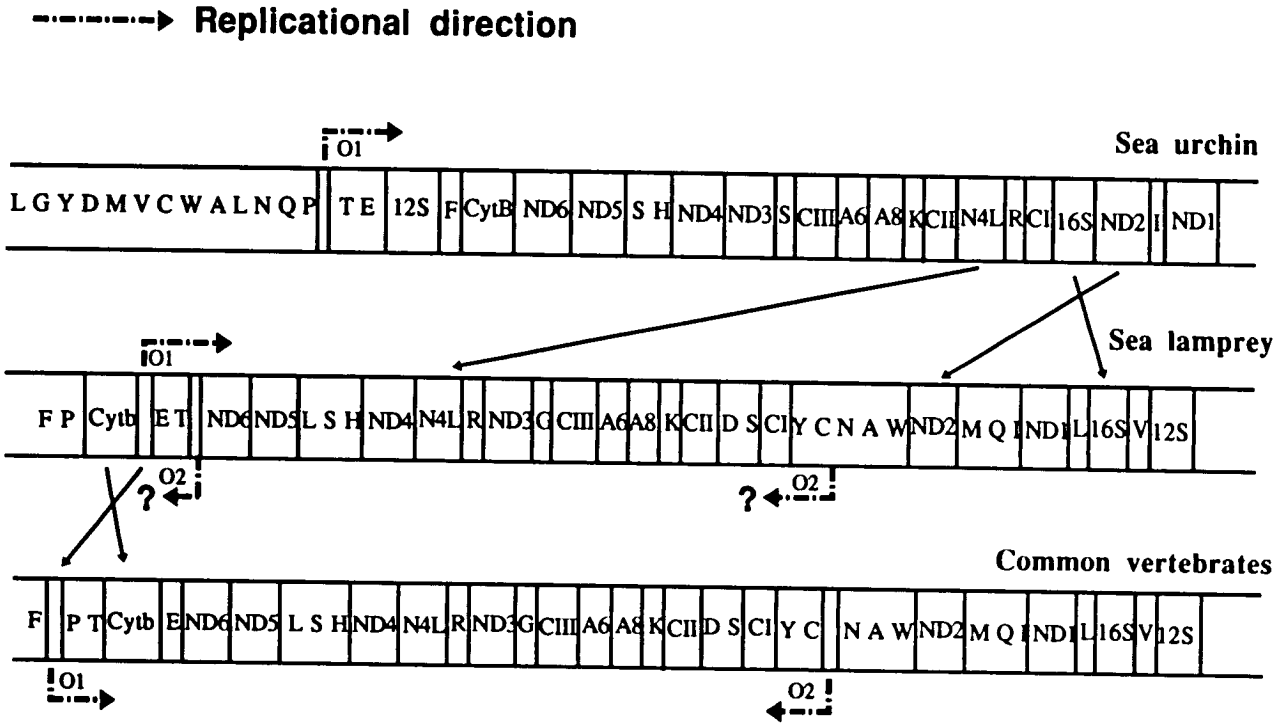


Figure 3-7. Comparisons of gene arrangements among three deuterostome mitochondrial genomes. The movements of tRNAs are not indicated. The gene order of protein genes in the three genomes is colinear except for ND4 L in sea urchin mtDNA. A and B in the lamprey genome represent the two major non-coding regions. The second non-coding region (B) nearly consists of seven copies of a tandem repeat. At this point, the location of the second strand replicational origin in lamprey mtDNA is not known. For more detail in echinidrms, see taxt.

order. Duplication of segments, followed by internal deletions, is probably a major mechanism. Duplications are frequently observed (Moritz et al., 1987), and seem to explain the unique order seen in birds. Cantatore et al. (1987) suggested that rearrangement of tRNA genes might also occur by illicit priming of replication. This mechanism would lead to the accumulation of tRNA genes in a cluster near the replication origin. Neither of these mechanisms explains the inversion of segments observed among echinoderm classes (Smith et al., 1993).

Jacobs et al. (1989) disputed the suggestion that the clustered tRNAs of echinoderms is a derived state. Clustered tRNAs are found in at least four echinoderm classes (Smith et al., 1993), and so is not unique to urchins. Furthermore, the leucine tRNA which Cantatore et al. (1987) suggest was recently inserted near the urchin replication origin, is found in an identical position in the tRNA cluster of sea stars. They suggest alternative mechanisms, in which tRNAs might have dispersed through the genome on the lineage leading to vertebrates.

The lamprey sequence demonstrates that the major rearrangements which distinguish echinoderm and vertebrate mitochondrial genomes did not occur recently on the vertebrate lineage. At this point, it is impossible to determine whether the lamprey gene order represents an ancestral or a derived state for the vertebrate lineage. Study of even more distantly related chordates may shed light on the sequence of rearrangements by which echinoderm and vertebrate mitochondrial gene orders have diverged.

Phylogenetic Significance

Often the comparative approach with DNA sequences from a single gene is not successful to resolve problems in the phylogeny of distantly related taxa, because the sequences accumulate changes so fast, which hides the real distances behind the superficial similarities. Because of the slow rate of gene rearrangements, the pattern of mitochondrial genomic organization may be informative of the topology of deep phylogenetic divergences (Brown 1985; Moritz et al, 1987). Each phylum has a basic pattern of gene arrangement. In echinoderms and insects, the gene order provide useful phylogenetic information at the class level (Fig. 3-8).

Even though the lamprey mtDNA sequence does not provide a definitive answer to the question of whether this novel gene order is an ancestral pattern to the other vertebrates, occurred in the agnathan lineage, or is unique to the sea lamprey lineage, it is clear that most features of vertebrate mt genomes well already established at an early stage of evolution. Thus, the features in common between lamprey and other vertebrates can be considered ancestral.

As more exceptions from the standard mtDNA structures are being revealed, more extensive investigations are desired to determine the usefulness of the pattern of gene rearrangements for the phylogenetic analysis of animals.

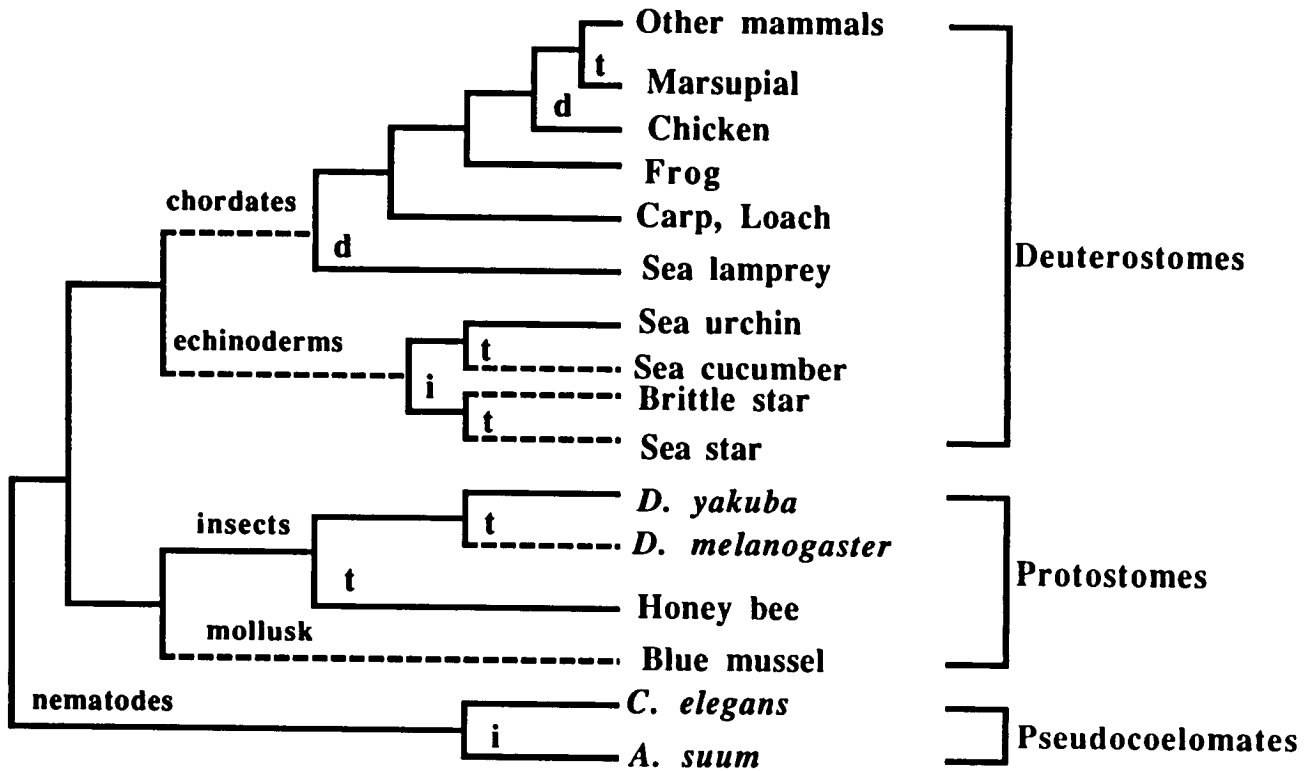


Figure 3-8. The pattern of mt gene arrangements among animal phyla. The t, d and i respectively represent translocations of tRNA genes, duplication/deletion events, and inversion within each phylum. Dashed (---) lines indicate partially characterized mitochondrial genomes.

Acknowledgement

I thank S. Sower for donation of lamprey tissues. I also thank W. K. Thomas and W. M. Brown for their helpful comments during the early stage of analysis.

CHAPTER IV

STRUCTURE OF TELEOST MITOCHONDRIAL CONTROL REGIONS AND EVOLUTION INFERRED BY COMPARISONS WITH LAMPREY AND MAMMALIAN CONTROL REGIONS

ABSTRACT

I amplified and sequenced the mitochondrial control region from 18 species representing 8 families of teleost fish. The length of this segment is highly variable among even closely related species due to the presence of tandemly repeated sequences and large insertions. The position of the repetitive sequences suggests that they arise during replication, both near the origin of replication, and at the site of termination of the D-loop strand. Many of the conserved sequences observed in mammals are also found among fish. However, in comparison with the lamprey mitochondrial control region, the overall sequence homology is very low and only some of the conserved sequence blocks (CSB-II, and III) exist in the lamprey control region, suggesting that the rest of the conserved sequence blocks including the central conserved domain, are recent inventions and that the two blocks may have the most important functional constraints in the vertebrate control region. Study of potential secondary structures of RNAs from the conserved regions provides little insight into the functional constraints on these regions. The variable structure of these control

regions suggests that particular care should be taken to identify the most appropriate segment for studies of intraspecific variation.

INTRODUCTION

The decline of fisheries around the world has created new interest in genetic tools for management of fish stocks (Gauldie, 1991). Mitochondrial DNA (mtDNA) has been widely used as a marker for evolutionary and population studies because of its compact size, nearly complete maternal inheritance, and fast evolutionary rate (Brown et al., 1979; Wilson et al., 1985). Mitochondrial genes, because of their maternal inheritance, are expected to provide a more sensitive tool for detecting population subdivision than nuclear genes (Birky et al., 1989). Observations of mtDNA gene diversity thus provide some of the best information for studying levels of gene flow among fish populations (Ovenden, 1990).

Numerous studies of mtDNA haplotype diversity in fish have been conducted using restriction fragment length polymorphisms (RFLPs) of the entire mitochondrial genome (Avise et al., 1988; Graves et al. 1992). Recently, the use of nucleotide sequence data, rather than RFLPs, has been encouraged, primarily because of the greater sensitivity of sequencing in detecting variants (Beckenbach, 1991; Bartlett and Davidson, 1991; Carr and Marshall 1991a,b; Finnerty and Block, 1992). Polymerase chain reaction techniques make it feasible to target particular gene segments carrying the highest density of intraspecific variation in large numbers of individuals (Kocher et al., 1989).

Many studies of mammalian mtDNA have focused on the major non-coding region, located between the tRNA-Pro and tRNA-Phe genes, because of its supposedly rapid rate of evolution (Hoelzel et al., 1991). This region, often called the control region, includes transcriptional promotors for both strands, the heavy strand replication origin, and the displacement or D-loop region (Clayton, 1982; Chang et al., 1986). Because of reduced functional constraints, some portions of the control region evolve much faster than the average mitochondrial sequence (Brown, 1985). The segments directly adjacent to the flanking tRNAs often show the highest rates of base substitutions and insertion/deletion events (Saccone et al., 1987).

Although large variations in the size of fish mitochondrial genomes are known (Billington and Hebert, 1991), the extent to which this represents expansion of unique control region sequences is poorly understood. The exact size of the control region is known for only a few fish species. The sequence of the white sturgeon control region is 976 bp (Buroker et al., 1990). Two species of cyprinid also exhibit relatively short control regions (896-927 bp; Tzeng et al., 1992; Chang et al. 1994). Salmonid control regions are somewhat longer (1128 bp; Shedlock et al., 1992).

Most size variation of the control region at the intrafamily level is caused by the tandem repeats. Moreover heteroplasmy in the control region have often been found (Mignotte et al., 1990; Brown, 1992). Although no definitive processes that give rise to the tandem repeats or heteroplasmy in the control region have been demonstrated, it has been suggested that the DNA turnover or slippage might occur

in the mitochondrial control region (Hoelzel, 1993).

The purpose of the research was to describe the structure and evolution of some fish mitochondrial control regions in order to provide a methodological basis for studying fish populations in the Gulf of Maine and the Caribbean sea. I present complete control region sequences from 8 species, and partial sequences from an additional ten taxa. I compare the structure of these fish control regions with those of sea lamprey and mammals in an attempt to understand the functional constraints and pattern of evolution of this region in vertebrates.

MATERIALS AND METHODS

DNA Samples

The samples used in this study are presented in Table 4-1. They include three clupeids (alewife, American shad, and Atlantic herring), four gadids (Atlantic cod, haddock, pollock, and tomcod), redfish, cunner, and four flounders (American plaice, greysole, yellowtail flounder and winter flounder), all collected from the Gulf of Maine. The Caribbean reef fish include two damselfish (bicolor dameselfish, 3-spot damselfish), two wrasses (bluehead wrasse, yellowhead wrasse), and one parrotfish (stoplight parrotfish). Dameselfishes and wrasses were collected from Teague Bay, St. Croix, and parrotfish was from Lameshure Bay, St. John. Total genomic DNA was prepared from 1-2g of frozen muscle tissue or ethanol preserved sample by proteinase K digestion at 37°C overnight followed by purification through phenol/chloroform extractions and ethanol precipitation (Sambrook et al., 1989). The ethanol preserved samples were vacuum dried prior to proteinase K digestion. The sequence of lamprey control region is from the complete mitochondrial sequence described in chapter III.

Amplification

For the Gulf of Maine fish, I amplified and obtained the sequence of a part of cytochrome b, tRNA-pro, and the first half of the D-loop with primer L15774 (Shields

Table 4-1. Traditional classification of the fish species in this study (after Carroll, 1988).

Osteichthyes; Actinopterygii;
Chondrostei;
Acipenseriformes
Acipenser transmontanus (white sturgeon)
Neopterygii; Teleostei;
Clupeomorpha;
Clupeiformes
Alosa pseudoharengus (alewife)
Alosa sapidissima (American shad)
Clupea h. harengus (Atlantic herring)
Euteleotei;
Salmoniformes
Oncorhynchus mykiss (rainbow trout)
Ostariophysi;
Cypriniformes
Cyprinus carpio (common carp)
Crossostoma lacustre (loach)

Neoteleosti;
Paracanthopterygii;
Gadiformes
Gadus morhua (Atlantic cod)
Microgadus tomcod (Atlantic tomcod)
Melanogrammus aeglefinus (haddock)
Pollachius virens (pollock)

(Table 1. continued)

Acanthopterygii; Percomorpha;

Perciformes

Cichlidae

Champsochromis spilorhynchus

Cichlasoma citrinellum (Midas cichlid)

Pomacentridae

Stegastes partitus (bicolor dameslfish)

Stegastes planifrons (three-spot dameselfish)

Scaridae

Sparisoma viride (stoplight parrotfish)

Labridae

Thalassoma bifasciatum (bluehead wrasse)

Halichoeres garnoti (yellowhead wrasse)

Tautogolabrus adspersus (cunner)

Pleuronectiformes

Glyptocephalus cynoglossus (grey sole)

Hipploglossoides platessoides (American plaice)

Pleuronectes ferruginea (yellowtail flounder)

Pleuronectes americanus (winter flounder)

Scorpaeniformes

Sebastes marinus (redfish)

and Kocher, 1991) and E (Table 4-2). I then designed a primer (A,Table 4-2) complementary to the tRNA-Pro, making it easy to amplify the control region. First, using A and D primers, the whole control regions of six species (Atlantic cod, haddock. pollock, tomcod, alewife, and American plaice; Figure 4-2) were amplified. The sizes of the amplified products were too long to be sequenced without sets of internal primers. Thus, after obtaining the sequence from the first half of the D-loop, I designed several primers complementary to the highly conserved central region which were used to amplify the second half of the control region.

For the Caribbean reef fish, I amplified the first half of the control region with a pair of primers (A, E). The sequences and the locations of the primers are shown in Table 4-2 and Figure 4-1 respectively.

PCR of partial or complete control regions was conducted according to standard protocols (Kocher et al., 1989). For amplification, the following reagents were added to each microtube: 1 µg genomic DNA, 5 µl 10x Taq buffer (0.67 M Tris pH 8.8, 0.02 M $MgCl_2$, 98 mM β-mercaptoethanol, and 0.1% Tween-20), 5 µl of both primers (10 µM), 1µl 2.5 mM each dNTPs and 0.33 µl 5 unit Taq DNA polymerase. Each sample was brought up to 50 µl with sdH_2O. The reactions were overlayed with 20-30 µl of mineral oil to prevent evaporation during thermal cycling. PCR conditions consisted of 25-40 thermal cycles of denaturation (93°C, 0.5 min), annealing (50-55°C, 1 min), and extension (72°C, 2-3 min). After thermal cycling, the effectiveness of amplification was checked on a 1-2% Nusieve agarose gel. In many cases a single

Table 4-2. List of primers designed for amplifying the control region of various fish families.

A:	5′-TTCCA CCTCT AACTC CCAAA GCTAG-3′
B:	5′-ACGCT GGAAA GAACG CCCGG CATGG-3′
C:	5′-TTGCA GTTTT GTCCA TCTCT TA-3′
D:	5′-GTCCA TCCTA ATATC TTCAG TA-3′
E:	5′-CCTGA AGTAG GAACC AGATG-3′
*F:	5′-CGTCG GATCC AGAGC CTACC ACAAG GTGAT T-3′
*G:	5′-CGTCG GATCC CATCT TCAGT GTTAT GCTT-3′
H:	5′-TACAA TGGCA TAATA GGTGG CA-3′
I:	5′-TAAGT ATAGA AGAGA CTACC G-3′
J:	5′-TTTGG TTCCT ATTTC AGGGC CA-3′
K:	5′-AGCTC AGCGC CAGAG CGCCG GTCTT GTAAA-3′
L:	5′-AGTAA GAGCC CACCA TCAGT-3′
M:	5′-TATGC TTTAG TTAAG GCTAC G-3′
N:	5′-GGGGC GCGGA TCCCA TCCTA ATATC TTCAG-3′
O:	5′-CAAGC CGGGC GTACA CTCCA G-3′
P:	5′-CCGGA AACAG GACAA ACC-3′
Q:	5′-GGGCG GATCC CACCA CTAGC TCCCA AA-3′

* *BamHI* sites (CGTCGGATCC) were added to these primers to facilitate cloning.

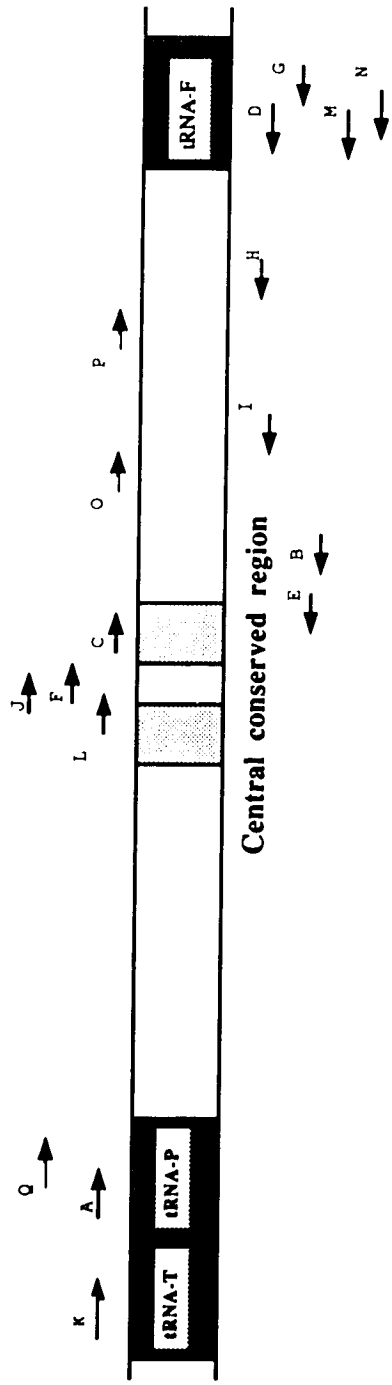
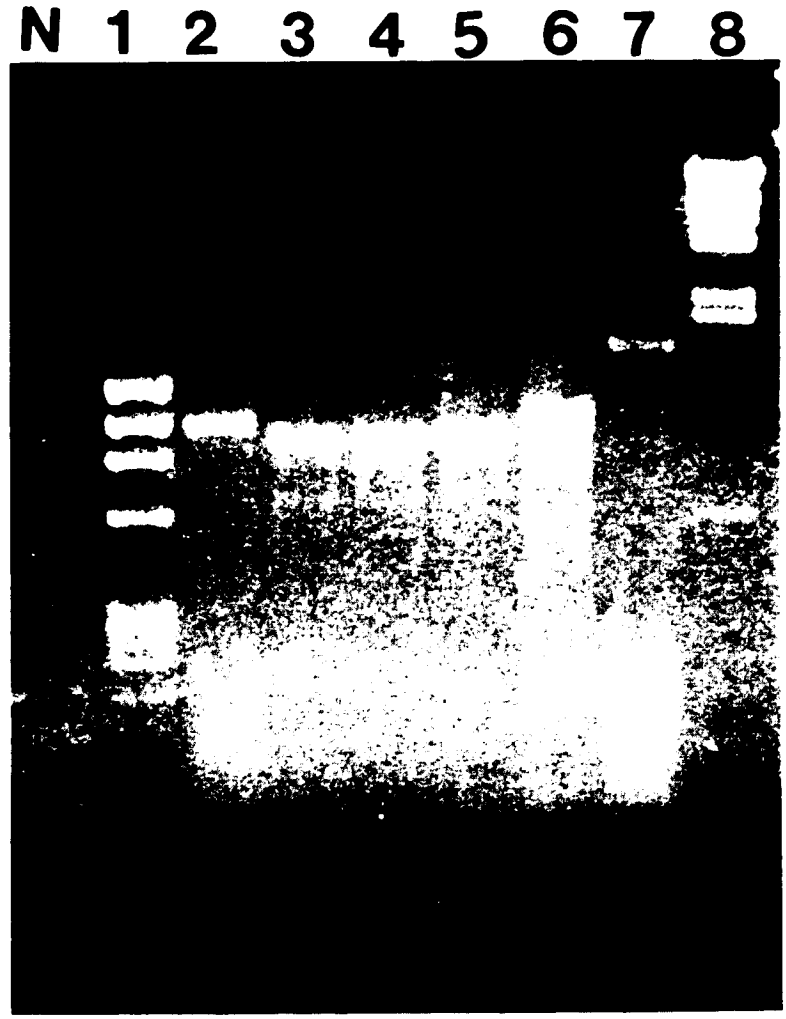


Figure 4-1. Sequencing strategies for the control region. Arrows indicate the directions of DNA synthesis from the primers. For the sequence of each primer see Table 4-2.

Figure 4-2. PCR products of the entire control region from six teleosts. Lanes 1 and 8 are DNA size markers, and lanes 2 through 7, represent Atlantic cod, haddock, pollock, alewife, and American plaice respectively. Among four gadids, the product from Atlantic cod is a little larger, because of the presence of a tandem repeat in the region flanking tRNA-Pro. American plaice contains the largest control region due to an expansion in the second half of the control region.



band was cut from the gel to avoid sequencing of primer dimers. Amplified DNA products were retrieved from the total PCR reaction by one of three methods: phenol/chloroform extraction after melting the gel band in TE buffer, gelase digestion (Epicentre Technologies) of the agarose slice, or Centricon-100 purification (Amicon Co.) directly from the PCR reaction.

Cloning

Although direct sequencing of the amplified products worked well for most samples, the second half of the control region of the three flounders were difficult to sequence because of poor amplification and the large size of the amplified product. For these samples I amplified using primers F and G (Table 4-2) containing a *Bam*HI restriction site. The PCR products were ligated into *pBluescriptII* plasmid digested with the same restriction enzyme and dephosphorylated with CIAP. Transformation of ligated plasmid and plating methods followed the manufacturer's (Stratagene) instruction manual. After identifying the recombinant clones, I obtained the double-stranded recombinant DNA using standard alkaline-lysis miniprep methods (Sambrook, 1989).

The Caribbean reef fishes were amplified well with primers A and E. However, for unknown reasons, the sequences contained a high level of background noise. Hence, the PCR products amplified with primers carrying *BamH I* restriction sites were cloned into M13 phage vectors (BRL), according to the manufacture

protocol. Single-stranded M13 recombinant DNA was rescued by alkaline-lysis followed by 13% PEG_{800} precipitation.

Sequencing

I sequenced amplified double-stranded DNA, double-stranded plasmid DNA, or single-stranded M13 DNA on an automated DNA sequencer (ABI 373A) using the Taq DyeDeoxy Terminator cycle sequencing kit (Rev.e; Applied Biosystems Inc.). Sequencing was performed with the amplification primers, additional internal primers where necessary, or the universal M13 primers (seeTable 4-2). After the cycle sequencing reactions, unincorporated dye terminators were removed with Centrisep spin columns (Princeton Separations. Inc.) and the sample vacuum dried. The entire sample was resuspended in 4 µl of formamide loading buffer, denatured and loaded on the sequencer.

Sequence Analysis

Analyses were performed using programs in the GCG package (Ver. 7.0, Genetics Computer Group Inc.). I searched for repetitive elements using the REPEAT program with 80% stringency. The base composition was calculated with COMPOSITION program. The sequence alignment of the mouse and fish sequences was performed using PILEUP (gap and gap length penalties used were 1.0 and 0.3 respectively). No obvious homologies between the mouse and the fish sequences were

identified with this approach. I therefore performed dotplot analyses with COMPARE and DOTPLOT programs to identify sequence features which are conserved in fish. Additionally, I compared each 20 base segment among four gadid sequences in an attempt to find the most conserved region. Conserved sequence blocks identified in this way were compared to conserved features of mammalian control regions. After alignment, potential secondary structures of the conserved segments were examined using the FOLD program. The final alignment was produced using the text editor ESEE (ver. 2; Cabot and Beckenbach, 1989). Labelling of conserved sequence blocks follows that of Southern et al. (1988).

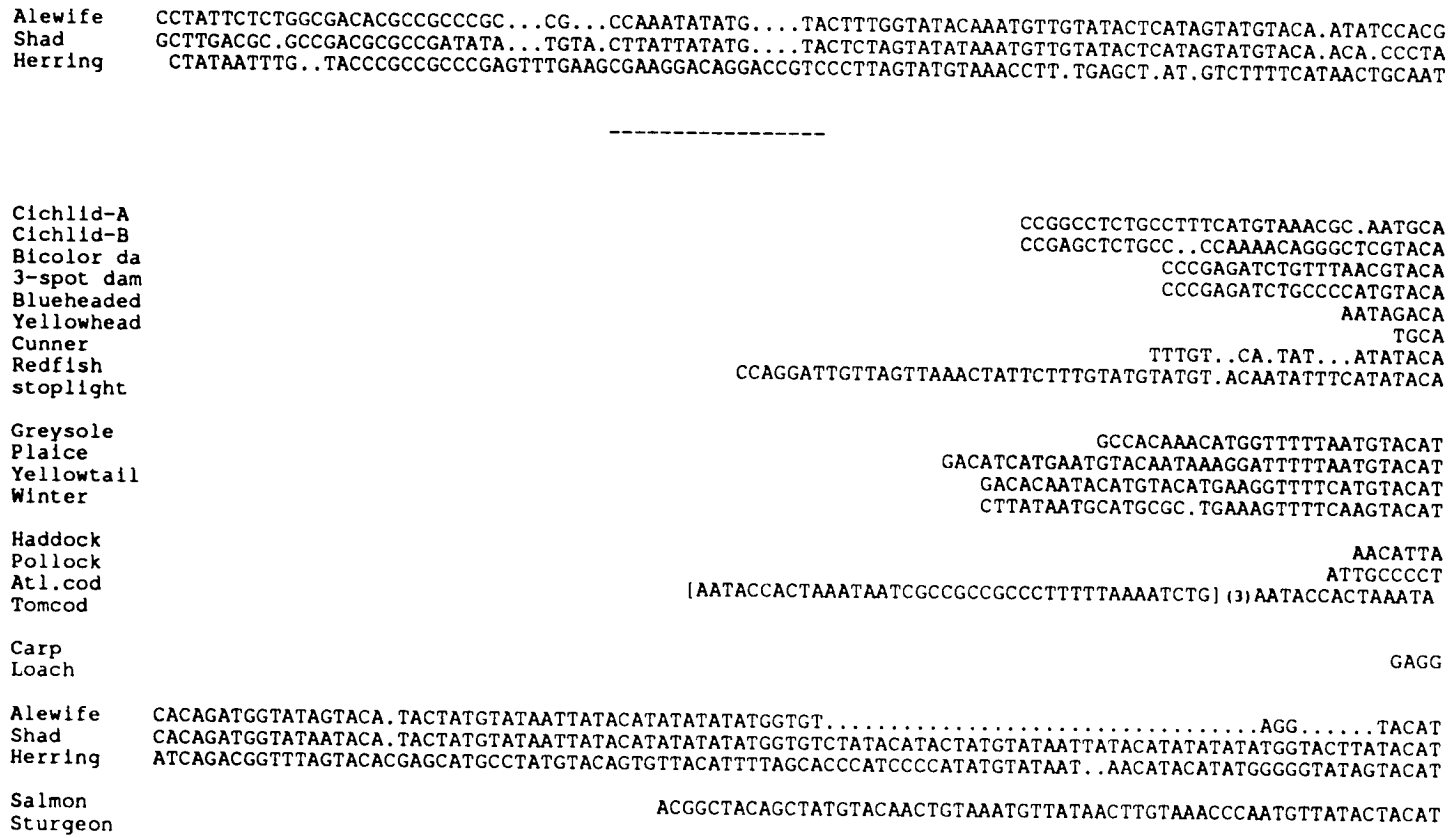
RESULTS

Alignment of Sequences

I first tried a PILEUP alignment of our fish sequences with that of the mouse (Bibb et al., 1981) to identify the location of conserved sequence blocks (CSBs). This alignment was not satisfactory in that mouse CSBs were not aligned with conserved segments of the fish sequence. As a next step, dotplots were constructed to identify sequence segments conserved among fish. The dotplot searches showed a conserved central region even at 95% stringency. The final alignment was produced manually, through an approach similar to the PILEUP algorithm. The sequences of each family were aligned first, paying attention to the central conserved segment, and other CSBs. Next, these sequences were aligned to other families, starting with the most conserved regions. This alignment, while not definitive, allows the identification of most of the conserved elements defined in mammalian sequences (Southern et al., 1988).

The ESEE alignments of the 23 new sequences, together with four previously reported fish sequences, are presented in Figure 4-3. The size of these control regions varies from 856 to 1500 bp. The longest control regions are found in the pleuronectids, due to expansion of the 3' portion of the sequence. Cichlids and gadids contain relatively short control regions, although they are still longer than the control

Figure 4-3. Aligned sequences of the mitochondrial control region of 18 fish species. The alignment contains 13 complete control region sequences and partial sequence (5'-region only) from cunner, redfish, greysole, shad and herring. Dots (.) denotes gaps inserted to maximize similarity among the sequences. The most conserved sequence block in fish, corresponding to mammalian CSB-D, is marked. The central conserved region used for phylogenetic analysis is also indicated. In the second half of the sequences from two flounders (winter flounder and yellowtail), the dashed lines (-) indicate regions for which sequence was not obtained, but which most likely contain additional repetitive sequence units. Sequences of the carp (Chang et al., 1994), loach (Tzeng et al., 1992), salmon (Shedlock et al., 1992), and sturgeon (Buroker, et al., 1990) were obtained from the Genbank. Cichlid-A is *Champsochromis spilorhynchus* and Cichlid-B is *Cichlasoma c.f. citrinellum.*



(Figure 4-3 continued)

```
Cichlid-A  TATATGTATTATCACCATTATTT.TATGTT.AApACATATCCTATAT.ATT..AATACATATCATTTACAAAAA.CATAGATA..AATA.TACCACATAT
Cichlid-B  TATATGTACTTACACCATGAATT.TATATT.AACCATTTATTAATG.ATAC..AAGACACACCATTAAAGGTGA.CATCCCTACTCTTACTAACACCCTC
Bicolor da TATATGTACTTACCCCATTAATT.TATATGCAACATTAAAAGTAG.GGCATTCAAGAACATAAGTGATTAATAAACTAAAATTTATTGAACCCATGGATA
3-spot dam TATATGTAATTACCCCATTAATT.TATATGCAACATTAAAAGTAGAGGCATTCGAGAGCATACATGATTAATCAACTAACTATTATTTTAAACATGCTTG
Blueheaded TATATGAACTTCTACCATAATAT.GGTATG.AACCATATCACTGTA.CTTTAACAGAACACATATGTATTACCAACATTTCTAGGAGTTAAACATTCATT
Yellowhead TATATGATTTTACACCATATTAT.GGTGTG.AACCATATAACTGTA.CCTCTTTGACACATATATGTATTATCAACATTTCTTGAATTATAGCATTCATA
Cunner     TATATGTAATTACACCATAAATT.TATATT.AACCATATCA..ATA.GTATTCTAGTACATCTATGTATAATAACCATATCTAGGGTTTAACCATTCAAG
Redfish    TATATGTATTATCACCATTAATT.TATATT.AACCATATCA.TAGG.GCATTCATATACATATATGTATTATCACCATATATAGAATTTAACCATTCAAG
stoplight                                                                                                ACTCAGACATC

Greysole   GTATGTAATA.ACACCATATATT.TATAGTACCCATTTTATGTAATG.....TACTAAGACA.TTCATGTATAATAACCCAATCTAGTAATAAAACACTC
Plaice     CTATGTAATA.ACACCATATATT.TATAGTAACCATATCGTATAGTG.....TACTAGGACA.TACATGTATAATAACCTAATCTAGTAAAATAGCACTC
Yellowtail GTTTGTAATA.ACACCATATATT.TATAGTAACCATTTTATATAATG.....AACTAGGACA.TTCATGTATAATAACCTAATCTAGTAATTAAGCACTC
Winter     GTATGTAATA.ACACCATATATT.TATAGTAACCATTTTATATAATG.....AACTAGGACA.TCTATGTATAATAACCTAATCTAGTATTATAGCACTC

Haddock    TTAGAGCCCCGCCCTTTTAAAATATAAATTATCGCGCGAGTTAACCCCACCTTTTTTTT.......TTCCTTTATGTACTTATGTCCTAGAAAACATAAC
Pollock    AGGGATCCTCCCCCTTTTAAAATATAAAT..CCGCGCGAGTTAACCCCACCTTTTTTTT.......TTGTGCTTTTCCTGCATGTCCTAAAAGCCCTATC
Atl.cod    ATCGCCGCCGCCCTTTTTAAAATATTGATTTTCGCTTGAGCTTACCCCTCCTTTTTTCTCCCCCCCCCCCTGCCTCATAATATGTCCTAGAGATCATATT
Tomcod     GATACCCCACCCCCTTTAAAATAT.AATTTCCGCGCGAGTTATCCCCACCCTTTTTTT........TTTA.CCATG.TACAGATGTCCTAGATATCTTAT

Carp       TAGCACTCCCTTTATGGTATAGTACATATTATGCATAATAT.TACATTAATGTATTAGTACA.TATATGTATTATCACC.AACTCACTATTTTAACCATA
Loach      AATGCTGCAGGTATGGTATAGTACATATTATGCATAATAT.................TACA.TATATGTAATATCACC.ATTAAATTATTTAGACCATA

Alewife    A.CTAT.GTATAATCCCCATTCATATTATGTCAGGTAAATAACTGCTT...TACA.TTACATAACTGAATCTAAGAACAATACAATAATAACCAAATAAC
Shad       A.CTAT.GTATAATCACCATTCATATTATATCAAGTAAATAACTGTTT...TATACTTACATTCATGAATCTAAGGATAAATAAGTAATAATCAGATAAT
Herring    ATGTATACTATTATACATGTACCTA.TGTATCAA.TACATGTGTGTTTAACTATACATATACCTATG.GTATTAATACATCTATGTATTATCACCATACG

Salmon     CTATGTATAATATTACATATTAT.GTATTTACCCATATATAATACTGCTTGTGAGTAGTACATTATATGTATTATCAACATACGGTGATTTTAACCCCTC
Sturgeon                                                                                    GACCATGCTATGTTTAATCCACATTAATTTCTAGCCACCATAC
```

(Figure 4-3 continued)

```
Cichlid-A  TTGTTTAAACCATTTTAACTA.AGGGGTACATAAACCATAACTGA.ATATACTCCAATAAACCT.TATTAACCACTGAACGATA.......GTTTAA.GA
Cichlid-B  ATCCATCAATTATTCCCCTAA.AATTTTACATAAA.CATAAAAAAGATA.AATCAATTAAACAT.GAATAA.GAATT.AACGAC.......ATTTAA.GA
Bicolor da TTAATCAAGAATTAAAATCTAAGGGTTACA.TAAGCATTAATTGAAATAAGCAAAATTATATTAACCTGAAACTAGACCGAGA.........TTTAA.GA
3-spot dam TTTCGCACATATAAAACTAAGACTCAAACG.AAAGCATTAAAATCAACGACCAAAA.AACAATGAAATAATACT.GAACGAAA.........CTTAA.GA
Blueheaded CATCAACATAACACCTAAGACTTACAAGAA.C...CATAACTGACCAAACAGAAGA..A...ACAGTTCTGAAGGCTGAGAGA.........TTTATTGA
Yellowhead CAACACCATTACA.CTAAGAGATACATAAA.G...CATTCAATGAATAATCCATAC..A...ACTGATATATCTACTGAGAGA.........TTTATTGA
Cunner     TATTACCCCAAAAGAAGAACTATGCATAGA.G...CAGTACTGGTATATTTAATATTTA...TAGAATTCAAGGATGGGGAGA.........TTTAA.GA
Redfish    TATTATACGACACGAA.CATTTTACATAAA.GTAAAATAATAAGACTAACAAACACTTA...TA.AATACCA....AGCGAAA.........TTTAA.GA
stoplight  AATCAAGCAACTGCAGTTGCCCAACTGACATCAAACCATAAAACTATGTAGTATACCCATAGTATAATGTAAAACCCGAGAGA.........CTTAAATA

Greysole   ATTTATCACCACTTTTAACTAAAATATACTACAACCTGAACGATTACTG.....ACCTTAAATTAGTGAAACTCCAGGACGGGCCGAA.AGACCAGACCG
Plaice     ATACAACACCA.TTATAACTAAAAAATACTAGAACCCAACTTACTACTA.....ATATAACATATGTGAAAGTCAAGGAGTCGTCGAG.ATTTAAGACCG
Yellowtail ATTCATCAACAATTTTACCTAAGATAGACTAGAACCTAGACAAGCACTA.....ACCTTACATTAGTGAAATTCCAGGACCAGTCGAA.ACTTAAGACCG
Winter     ATTCATCAACATTTCTAACTAAGATTTACTGAAACCTATTTTTACACTA.....ATCTTACACATCTGAAAGTCCAGGACCAGTCGAA.ATTTAAGACCG

Haddock    ATGTGTTTAAA..........GTACATATGTATAATCGCCATTAATTAACTTAACCATTCAAGGAGAAATAATCATGAAAATTTAACCATTCAAGCGAA
Pollock    TTTTTAGTAAAATTATTATGA.GTACATATGTATAATCACCATTAATTAACTTAACCATACAAGGAGAAATAATCATGACAAACG.ACCATTCAGGTGAA
Atl.cod    CTTTTTAGTAAATTATTATTT.ATACATATGTATAATCACCATTAATTAAGTTAACCATACAAGGAGAAATAATCATGAATAGTCAAGCATTCAGGATAA
Tomcod     TCTTTTGGGTAAGTCATTTAA.ATACATATGTATAATCACCATTAATTAACTTAACCAATCAAGGAAGAATAATTATAAAAACCC.ACCACTCAAGAT.A

Carp       AAGCAGGTACA.TAATATTAAGGTGGGCATAAAGCATATCATTAAGA.CTCACAAATTCTATTA.TTTGAACTTGAGTAATATATTAATCCCCAAAAATT
Loach      AAGCAGGTACATTACTATGTATGTAG..AATGAGCATAAATTTGAGAACTCCCATATATTTT...TTTAGACCTGGGTAATCGAATAATCCCCATAAAGC

Alewife    AATAAATAAGGACACAGCAAGTAA.TAATTGAACTAAGGTATACA.TAAGCATTAAATTAAGATTC.AGAATATAAATAAGAAAACCTGATAAATAGATT
Shad       AACCCATGAAAACA.AACAAGTAA.TAATTAAACTAAGGTATATACAAAGCATTAAACTAAGATTC.AGAATATCAATAAGAAAAGCTGATAAATAGATT
Herring    CTGAAATTA..CCCTATCAGGAAATTGATTAAATGAAGAAATACATCGA.CATAATTATAA.ATACGAGCCCATTATTAAGATCACCTGATAAATAGAAT

Salmon     ATACATCAGCAC..AAATCCAAGGTTTACATTAAGCCAAACACGTGATA.....A...TAACCAACTAAGTTGTTTTAAACTGATTAATTGCTATATCAA
Sturgeon   CATAATGCTCACAAGCACATTAAATTGTTTAAGTACATAAGACATGCTATGTTTAATCCACATTAATTTCTAGCCACCATACCATAATGTTTCATCTACC
```

(Figure 4-3 continued)

```
                                                                                                |→ Central conserved region
Cichlid-A    CCGATCACAACT.CTCACCGGTTAAGTTATACCAAGTA.CCCACCATCC.TATTCATTCCCCATATTTA....ATGTAGTAAGAGCCCACCATCA.GTTG
Cichlid-B    CCGAACACAAATAACCACTAGTTAAGTTATACCAAGTA.CCCAACATCCCG.TCAGTTCCCCCATCTTA....ATGTAGTAAGAGCCCACCATCA.GTTG
Bicolor da   CCGAGCACTTACTTCTCATCTGTCAAGATATACCACGAATCCAACATCCCTTATAAATACCCAATATTTA...ATGTAGTAAGAACCGACCATCA.GTTG
3-spot dam   CCGAACACTTACTTTTCATCAGTCTAGATACACCACGAACCCAACATCCCGTCAAACAGTCCAATATTTA...ATGTAGTAAGAACCGACCATCA.GTTG
Blueheaded   CCTAACACTCCCGTC.CGCGAGTCATGCCACACCAAGTCTCCCACTACCATAGTTTTAAACAGTTATGCG......CAGTAAGAACCGACCAACAAGTCC
Yellowhead   CCCAGCAAACCCGTC.CACCTCCAAAGACAAACCAAGTCCTCCACCCCTCTGATTTCAAACAGTAATGCG......CAGTAAGAACCGACCATCCAGCTG
Cunner       CCTAACACAAA.TCC.CACTGGTTAAGTTATACCAAGTAATCCACATCCCGCTCTTA.GAAAATTCTTA....ATGTAGTAAGAGCCGACCAACAAGCAC
Redfish      CCGAACACAAACACT.CATAAGTTAAGTTATACC.TTTACTCAAAATCC.TATCAAACTCAAATATTTA....ATGTAGTAAGAGCCGACCAACAAGTCC
stoplight    ATCACCATATCCGTC.CACAGCAAAAGATATGCCGAGT...AAAATACATCTCTATTAATATAAAGTTATAATGCACAGTAAGAACCTACCAACAAGCAC

Greysole     ATCACAACACTC.ATCAGTCGAGTTATACCAGG.ACTCAAGAT.CTCGCCCTCCCGAAAATTCT.........ATGTAGTAAGAGCCTACCAACCGGTGA
Plaice       AACACAACACTC.ATCAGTCGAGTAATACCAAG.ACTCAAAAT.CCCTTCAACATCAAAACCCT.........ATGTAGTAAGAGCCTACCAACCGGTGA
Yellowtail   AACACAACACTC.ATCAGTCGAGTTATACCAAG.ACTCAAAAT.CTCTTC.ACGCCAAAATTCG.........ATGTAGTAAGAGTCTACCAACCGGTGA
Winter       AACACAACACTC.ATCGGTCGAGTTATACCAAG.ACTCAAAAT.CTCTCC.AACTCAAATTCGT.........ATGTAGTAAGAGCCTACCAACCGGTGA

Haddock      AATAACAATGAATTAATAGAGCAAATATGGTTATTGTAACCAATTTATGGAATTTCGTACAAGAAA.......TTGTAA.ACATAACCGGACTTTCCTTG
Pollock      AACAACAATAGATTAATAGAACAAATATGGTTTTTTAAACCAATTTATGGATTTTTATACAAGAAA.......TTGTAA.ACATAACCGGACTTTCCTTG
Atl.cod      AACAACAATATATTTATAGGACAAATATGGTTATTTTAACCAATTTATGGAATTTAGTGCAAGAAA.......TTGTAA.ACATAACCGGACTTTCCTTG
Tomcod       AACAACAATAGATTAATAGAACTAATATGGTTATTTTAAACAATTTATGAAATTTGGTGCAAGAAA.......TTGTAA.ACACCAACGGACTTTCCTTC

Carp         TGTCCTCAAATTTTTCCTTGAAATAATCAACTATAATTTTATTCAAACATATTA...................ATGTAGGTAAGAGACCACCAACCAGTT
Loach        CCATCATAACATTTTCTTTGAATAAATTTGCCCACATCTTCTGAGTAATAAATA...................ATGTA.GTAAGAAACCACCAACCAGTT

Alewife      AATCCCTATTACTCCATTAAACCATTTTCCTTGCGTTACCCATCAAAAATAGCTA..TATACTTATTTA....ATGTAGTGAGAACCGACCAACACGACT
Shad         AATCCCTATTACTTCATTAAACCATTTTCCATGCGTTACCCACCAAAAATAGCTA..TAGACTTATTTA....ATGTAGTGAGAACCGACCAACACGACT
Herring      AATCCCCATAACTT.......CCAGTTTCCATGCGTTACTCAATATTAATCGGTAGCTCAAACGATT......ACCCAATAAGAACCGACCAACCAAATT

Salmon       TAAAACTCCAA..TTAACACGGGCTCCGTCTTTACCCACCAACTTTCAGCATCAGTCCGGCTTA.........ATGTAGTAAGAACCGACCAACGATTTA
Sturgeon     ATTAAATGTTA..TACACCATTATTTTTATGTGCACTAACATGA.TAAGCCTCCGATAACTTAAA........ATGTAGTAAGAGCCGAACATGGAGATA
```

(Figure 4-3 continued)

```
                                                          [GTGGG-box]                          [        CSB-D        ]
Cichlid-A  ATTCCTAA..ATGTTAACGGTTCTTGAAGGTCAAGGACAAGTATTC.GTGGGGGTTTCACTAGGTGAATT.ATTCCTGGCATCTGGTTCCTATTT.CAGG
Cichlid-B  ATTCCTTA..ATGATCACGGTTCTTGAAGGTGAGGGACAATTAT.C.GTGGGGGTTTCACTTAGNGAATT.ATTCCT...
Bicolor da ATTTCTTA..ATGCATACTCTTAATGATAGTGAGGGACAAGAACT..GTGGGGGTTTCACTTAG.TGATCTATTCCTGGCATTTGGTTCCTATTT.CAGG
3-spot dam ATTACTGA..ATGCATACTCTTATTGATAATCAAGGACAGTTATT..GTGGGGGTTTCACTTAG.TGATCTATTCCTGA
Blueheaded ATATCTTA..ATGCATACTCTTATTGAAGGTGAGGGACAAGTATTT.GTGGGGGTAGCACCTCAATGAACTATTCCTGG
Yellowhead A...AATA..GCGCATACTCTTATTGAAGGTGAGGGACAATAATT..GTGGGGGTAACACCTAG.TGAATTATTCCTGG
Cunner     ATTTCTTA..ATGCCAACGGTTATTGAAGGTGAGGGACAAGTATTT.GTGGGGGTTTCACCTTG.TGAACT...
Redfish    ATTTCTTA..ATGTCAACGGTTATTGAAGGTGAGGGACAATTATT..GTGGGGGTTTCACACAG.TGATTTATTCCTGGCATCTGG...
stoplight  ATATCTTA..AAGCATACGTTTCTTGATGGTCAGGGACAGTTCA...ATGGGGAGTAGTTTAACTTGCACTATTCCTGGCATCTGGTTCTACTT..CAGG


Greysole   TTCCTTA...ATGATAACTC.TTATTGAGGGTGAGGGGCAAAAATC.GTGCGGGTTTCACTCAGTGAACTATTCCT...
Plaice     TTTCTTAT..ATGATAACGG.TTATTGAAGGTGAGGGACAAAAATT.GTGGGGGTTTCACTCGGTGAATTATTCCTGGCATTTGGTTCCTACTT.CAGGG
Yellowtail TTTCTAA...ATGATAACGG.CTATTGAAGGTGAGGGACAAAAATT.GTGGGGGTTTCACTCCGTGATTTATTCCTGGCATATGGTTCCTACTT.CAGGG
Winter     TTCCTTA...ATGATAACGG.TTATTGAAGGTGAGGGACAAAAATT.GTGGGGGTTTCACTCGGTGAACTATTCCTGGCATATGGTTCCTACTT.CAGGG


Haddock    CC.....AAGGTAA..ACTGTCCAATGAAGGTGAGGAGCCCACATAGAAGACCCACCATCCCGTAACACGTTTCCTGGCTATTCTG.CCTAGCTTCAGGT
Pollock    CT.....AAGGCAA..ACTGTCCAATGAAGGTGAGGGGCTTAGACAGAAGACCCACCATTCCGTAACACGTTTCCTGGCTATTCTG.CCTAGCTTCAGGT
Atl.cod    CT.....AAGGCAA..ACTGTCCAATGAAGGTGAGGGACACATATTGAAGA.CCTCCATTCCGTAACACGTTTCCTGGCTATTCTG.CCTAGCTTCAGGT
Tomcod     CA.....AAGGCATA.TATACCCAATGAAGGTGAGGGACTTAAATTGAAGATTCACCATTCCGTAACACGTTTCCTGGCTATTCTG.CCTAGCTTCAGGT


Carp       TATATA.AAGGCATATCATGAATGATAGAATCAAGGACAATAA.TT.GTGGGGGT.ACACAATATGAACTATTACTGGCATCTGGTTCCTCATTTCAGGT
Loach      TATATA.AAGGTTAATTCTGCATGATAGTGTCAAGGACAAAAA.TT.GTGAGGGTAACACTTAGTGAACTATTACTGGCATCTGGTTCCTATTT.CAGGA


Alewife    AAAT...CGTGCATACTCTTAATGATAAGATCACGGACCCAAA.TT.GTGGGGGTTTCACAGAATGAACTATTCCTGGCATTTGGTTCCTATTT.CAGGG
Shad       AAAT...CGCGCATACGGTTAATGATAAGATCACGGACAAAAA.TT.GTGGGGGTTTCACAGAATGAACTATTCCTGGCATTTGGTTCCTATTT.CAGGG
Herring    GAGTT..AAGGCATATCATGAATGATAAGGGTCAGGGACACAAATC.GTGGGGGTTACACAGAATGAATTATTACTGGCATCTGGTTCCTATTT.CAGGG


Salmon     TCGGT...AGGCATACTCTTATTGATGGTCAGGGGCAGATATCGTATTAGGTCGCATCTCGTGAATTATTCCTGGCATTTGGTTCCTAAGT.CAAGGGCT
Sturgeon   TGTCT..AGAACATAAAGTTAATGA.GATGAGGGACAATAACTGTAGGGATTCACA...ACTGAACTATTACTGGCATCTGGTTCCTATTT.CAGGTCCA
```

(Figure 4-3 continued)

```
                                                  Central conserved region ←|
Cichlid-A  TCCTATACTTGTTATAATTCCTCATACTTTCATCGACGCTTGCATAAG.TTAATGGTGGTAATACATACT....CCTCGTTACCCACCAT.GCCGAGCGT
Bicolor da GCCATTGATTGTTATAATTCCTCACTCTTTTATCGACGCTTACATTGACT.AATGTTTGCCACCATTAAACTCGTT
Stoplight  GG.ATCCTCTAGAGTCGACTGCAGGCATG

Plaice     CCACGGATCGATA.TATTCCCCGCACTTTCACTAACACTTACATAAG.TTAATGTTGATAATACATACG....ACTCGTTACCCAGCAA.GCCGGGCGTT
Yellowtail TCATATACCGAT.TTATTCCCCACACTTTCATCGACACTTACATAAG.TTAATGTCGATAATACATACG....ACTCG.GACCCAGCAA.GCCGGGCGTA
Winter     CCATAAATTGATATTATTCCCCACACTTTCATCGACGCTTACATAAG.TTAATGTTGATAATACATACG....ACTCGTTACCCAGCAA.GCCGGGCGTA

Haddock    CCATCACAAGTAATTCGCTCACAAATTGCACTTTTGTC...CATCTC.TTAATGTATATACACATTATCTACTATAATCACTCCCCATGGCGGGCGTTCT
Pollock    CCATTGCTTGACCATCGCTCAAGATTTGCACTTTTGTC...CATCTC.TTAATGTCTATACACATTATCTACTATAATCACTCCCCATGGCGGGCGTTCT
Atl.cod    CCATTGCTTGACAATCGCTCATAATTTGCACTTTTGTA...CATCTC.TTAATGTATATACACATTATCTACTATAATCACTCCCCATGGCGGGCGTTCT
Tomcod     CCATTTGTAGGTAGTCGCTCAAAACTTGCACTTTTGTC...CATCTC.TTAATGTATATACACATTATCTACTATAATCACTCCCCATGCCGAGCGTTCT

Carp       ACA.TA.ACTGTAATATTCCACCCTCTTATAATTATACTGGCATCTGATTAATGGTG.TAGTACATACGT...TGCATTACCCCCCATGCCGAGCATTCT
Loach      ACATTATATTAAATATTCC.CCTAATTATTATTTATACTGGCATAT..TTGATGGTG.TAGTGCATACG...ACTCGTTACCCACCATGCCGAGCATTCT

Alewife    CCATAAATTTATAATCCCCCTAAGAATTGAACTTTC.CAGGCATAAG.TTAATGGTGGCATACTAA.CG...ACTCGTTACCCACCATGCCGGGCGTTCA
Shad       CCATAAATTTATAATCCCCCTTAGAATTGAATTTTC.CAGGCATAAG.TTA...
Herring    CCACACAGTCCTTTTCCCTCCCCTGGTGAACTATAT.CTTACATAGG.TTAATGGTAGTGTACTAAGG....GTTCTTTACCCAAGGC...

Salmon     ATCCTTAAGAAACCAGCCCCTGAAAGCCGAATGTAAAGCATCTGG..AATGGTGTCAATCTTA.TT..GCCCGTTACCCACCAA.GCCGGGCGTTCTCTT
Sturgeon   TTAACAGTTAT..TTCCCCATAACTGAACTATGTCTGGCATCTGA..AATGTTGGAGGTACTA.TGAAATCCATGACCCCACAT.GCCGAGAATCTTGTC

                                                  ----------------------

                                        [Py-tract]
Cichlid-A  TCTTTGCCAGCGTGTGGGGGGTTCTCTTTTT.TTTTNCCCTTTCATTTGACATCTCAGAGTGCATACAGAAATGACAGA........CAAGGTTGAACAT

Plaice     CACTCCAGCGGGTAAGGGG.TTCTCTTTTT..TTTTTCCCTTCCAATCGAATTAAAGGGTGAATCCAGCCA..TTAGA.AACGTCCAAGGGGTGAGAACT
Yellowtail CACTCCAGCGGCTAAGGGG.TTCTCTTTTTGTTTTTTCCTTTCACTTGGCATTACAGAGTGCATCCAGCCA..ACCGA.AACGTTCAAAGGGTGAGCACT
Winter     CACTCCAGCGGGTAAGGGG.TTCTCTTTTTG.TTTTTCCTTTCACTTGGCATTTCAGAGTGCATCCAGCCA..ACCGA.AACGTTTAAAGGGTGAGCATT

Haddock    TTCTAATGGACTACGGG.TTTCTTTTTTCTCCTCCACGTCAGTTGACATATCACAAGAGCAGAGCGTCGCCGCCGGA..........CAAGGTGGAGCTA
Pollock    TTCTAATGGACTACGGG.TTTCTTTTTTCTCCCCAAAGTCAGTGGAAATACCAGAAGAGAAGGGCGTCGCCGCCGAA..........AAAGGGGGAGCTA
Atl.cod    TTCTAATGGACTACGGG.TTTCTTTTTTCTTCTCCACGTCAGTTGACATATCAGCAGAGCAGAGCGTCGACGCCGGA..........CAAGGTGGAGCTA
Tomcod     TTCTAATGGACTACGGG.TTTCTTTTTTTCTCCTCAACTCAGTGGGAATACCAGAAGAGCAGAGCGTCGATGCCGAA..........CAAGGTGGTGCTA

Carp       TTTA.TATGCATA.GGG.TATCTTTTATTGGTTTCCTTTCATCTGGCATTTCAGAGTGCAGGCTCAAATGTTAAAT........TAAGGTTGAACATTTT
Loach      CTTAA.AGGCATA.AAG.TTACTTTTTTCCGATCACTTTCATAT.ACATCCCGGAATGCA.GCGCAATTGTATAAT........TAAGGTTGAACA..TA

Alewife    CTATACATGCATCT.GG.TTCTTTTATTGTGTCCTTAACTTTCACTAGGAATTTGGGGACTCCTCCCTAATGTTAGCGTG....TAGGGGGTGTACATTT

Salmon     ATATGCA..TAGGG.TTCTCTTTTTTTTTTCTTTTCCTTTCAGCTTGCATATACAAGTGCAAGCAAAGAAGTCTA......ACAAGGTCGAACTAGATCT
Sturgeon   AACATTTGGTATTTTTTATTCGGGTTTCCATTCACTGACATGTAGAACTCCTTCAGAGAAGAACAACAAGGTGGAACAT..ACGCGACTGTCCGAGAGGA
```

(Figure 4-3 continued)

(Figure 4-3 continued)

```
                  [   CSB-III   ]
Cichlid-A  TGT.AAACCCCCCGGAAA.CAGGA.AAAGCTCTAGAAGTGTTTT..TCGCACTCGTAACGTACAGACATGATTGTTATTCATAAATTGTTTGATGTGTAT

Plaice     GA.AAACCCCCCGGAAA.CAGGACAAACCTCTGGTAGCTTAGTAAAGGTGATTAC..TTTCAACCC.AAAACATTGGTAGTCTTACCCCCAGGCATACCA
Yellowtail A..AAACCCCCCGGAAA.CAGGACAAACCGCTGGCAGCTTAG.AAAGATGGCTAC[ATTCCAACCCTAAAATAACGGTGACCT.ACCCCA.GACTTTCCA
Winter     AGGCGGC.....................................[TACTTTTTCAACCCTAAAATAGCGGTAGCCC.ACCCCAAG.CTTACCG

Haddock    C..AAACCCCCCGGGAA.CAGGGAAATCCCTAGGACTGAACATATTTTGGTAGGAAATAACTAATAATGTTATAAAATTTGTTTTATTACATTATTTCCA
Pollock    C..CAACCCCCCGGAAA.CAGGGAAATCCCTAGGACTGGACATATTTTACCCAAATGTAACTAATAATGTTAT.AAATTTGTTTTATTTCATTATTTCCA
Atl.cod    T..AAACCCCCCGGAAA.CAGGAAAATCCCTAGAACTGAGTTTTTTTTATCAAAATATGACCAATAATG.TATAAATTTGTTGTTATTGCATTATTGCAA
Tomcod     C..CAACCCCCCGGAAA.CAGGGAAACCCCGAGGACTGGACATATTTTACCCAAATGTAAC.AATAATGTAATAAAATTTGTTTTATTACATTATTTCCA

Carp       CAAACCCC...GAAACCAAGGAGGACCAAGAACGTGT....GAACCAACGAGTTGAGGTATAAATTGGC..ATCCCATTATATATATATATATATATCGA
Loach      CAAACCCC...GAAACCAAGGAGGGCCTGCCGGGTGTACCAAAGCCAGCAAGTTGTGATAGGGGCCGGCTTATGCCACCGCATGTATATATATATAT.GA

Alewife    .AAACCCC...GAAACCATGAAAGACTCGACTGGCGTCATCAACGAGTTCTGTTACGTGTTGGTATATATAGTGTTGCAAAAAGATGTTACTGTGTGGCT

Salmon     ACCCC...TAAACCAGGAAGTCTCAAATCAGCAATATTTTTTT..ATATACATTAATAAACTTTTATGCACTTTAGCATTTGGCACCGACAGCGCTGTAA
Sturgeon   ACCCC...AAAAGCAGGA.CTGACTTGTCATCGACATACCTTGATC

                                                       ---------------------------

Cichlid-A  ATTATACTATTGCAATATTGCACTT

Plaice     TATGTTGAAGTATTATAATAATGCAATAATGCTTAAATTTTTTCCGCTAAAACAC...TAGCGGTCCGCCCCAAACTTACCAGCTATTGTTAAAATATAA
Yellowtail ATTATTGTAATAATGTTAAT](n).....A.TTC..CAACCCTAAAATAACGGTGACCTACCCCCCCCGAGCACT.....................
Winter     ATTATTGTAATAATGT](n).....ACTTTTTCAACCCTAAAATAGCGGTAGCCCACCCCCCCGGCAATT.....................

Haddock    ATTATTAAAATTT
Pollock    ATTAATAAAATTT
Atl.cod    ATTATTAAAATTT
Tomcod     ATTAATAAAACT

Carp       TCGGTTTTTTTAACCGCAACTTACCACTTA....CCTAAAAGTCCCTACCAAAAATCCCCAAAAAGAGGCTCGACACTAAATACTCTAATATAATTAATC
Loach      TACATAAACAAATATTAAACTTATATTTTAAAAGCCTAAAATTTAGTATTAAAAA.AGTATTAAATAACCTCAATACTAATATTTCAAAGTTATATT

Alewife    AGTGTAGC

Salmon     TGCGTACACTTCCATAAATAAAGTATACATTAATAAACTTTTCGATCCACTTTGTAGCACCTAGCACCAACAACGCTGTTATCAATGCCATTTCCACGCA
```

(Figure 4-3 continued)

```
Plaice      CTCTGACCCAAAACACCGGTAGTCTCTCCTAAACTTATTACCGATAGAATATCATAATATTACAATAATGTCAATATTTTTTACCATGGGCACAACAACC
Yellowtail  ....................................................................................................
Winter      ....................................................................................................

Carp        AGCTAGCGTAGCTTAACACAAAGCATAGCACTGAAGATGCTAAGATGAGTCCTAAGAAACTCCGCATGCA

Salmon      CAGCCCGCCGCTGACGTA

Plaice      CGCCCGAGACTTACCAGGGAAATAATATTGCAATATGCCAATATTTTTTCCCTGGAGCACCGACAACCCGCCCCAGACTTGCTACATGTTACGTATGAAT
Yellowtail  ....................................................................................................
Winter      ....................................................................................................

                                                    ----------------------

Plaice      AACTTTAAGAGAAAACCTCGGTAGTTTTTTTAAACCTGACCACCTATTATGCCATTGTAACATTATAGTAGTATCAACATGCCTTTCCTAAAACACCAAC
Yellowtail  ....................................................................................................
Winter      ....................................................................................................

                                                    ----------------------

Plaice      AGCCCACCCCAGACTTAGCAACAATCGTAAAGTAGGAACTTTCAACTCGAAGCACCGAGAGTTTCCTCAAGACTAACCAACTATTGTCTTCTTCATGGTC
Yellowtail  ..........................................................................CCAACTATTAGCCTCTTCATGATC
Winter      ..........................................................................CAAGTATTGTTTTCTTCATGATCAAAA

                                                    ----------------------

Plaice      AAAATACCATTATTTAAATTATTTCAAGTATTTTTGATATTAGAATTACCACGACCG.TGCAAAGTAAGAACCCGGGGAACGCTACCCCGTACCCAAAAT
Yellowtail  AAAATACTATTATTTAGATTATTTCAAGTATTTCCAATATTACCTCAACTTCGAAGGATGCAAAGCAAAACCTCAAAGAACCGTACTACTAATCGACAAA
Winter      TACTGTTATTTTAATTATTTCAAGTATTTCCAATGTTACTCCAACTACCAGAGACACAAAACGAAAACCCAAAAAACCTGAGCCCGCTGAAACTCATCTC

                                                    ----------------------

Plaice      TATTTCATTTTGACACAACACCAAGACT.............................CCT
Yellowtail  ACAAGCTCCTACTGAAGCTAGGCCCCCGTAGGTGGGAGACATCAGTAAAGTATTTACCT
Winter      CCC.............................AGGCCTCACTGGGGTATTTACCT
```

region of the sturgeon. Variation in size is mainly due to large insertions of unique sequence, although tandemly repeated sequences are found in several species. The unique features of each sequence are discussed by family in the paragraphs which follow.

Clupeidae

Primer pairs (Q,B) and (J,N) amplified unique products from each of the clupeid species. The overall length of the clupeid control region is approximately 1000 bp. All three species contained an AT-rich insertion adjacent to the tRNA-Pro. This segment was approximately 150 bp long in herring and shad, but only about 100 bp in alewife due to several indels. Clupeid control regions do contain central conserved sequences homologous to those of other vertebrates. The sequence of conserved block D (CSB-D) shows the greatest similarity to other species.

Gadidae

Primer pairs (A,B) and (C,D) generated unique products for each of the gadid species examined. Pollock, haddock, and tomcod have sequences of 868, 856, and 853 bp respectively. The presence of several copies of a 40 bp repeat near the tRNA-Pro in Atlantic cod makes its control region somewhat longer. Our sequence for Atlantic cod contains four copies of the repeat, and differs from the published sequence of an eastern Atlantic cod (Johansen et al. 1990) at 18 sites. These

differences include six transitions, 11 transversions and one indel. The first three copies of the repeat were identical and contained the sequence

5'-AATACCACTAAATAATCGAAGCCGCCCTTTTTAAAATCTG-3'

The fourth copy of the repeat is identical except for the last three nucleotides, which are more similar to the gadid consensus for this conserved region. High levels of sequence similarity are observed throughout the control regions of these gadids especially in the central conserved region (Fig.4-4). Gadids do not contain the GTGGG-box which is common in other euteleosts, and have a different consensus for CSB-D. The promoter-containing region between CSB-III and tRNA-Phe is very short.

Pleuronectidae

Primer pairs (A,E) and (F,G) gave unique bands for each of the flounders. The products from primers (F,G) were large and not easily sequenced. After cloning, sequence was obtained using four internal primers (H,I,O and P). The most striking finding is that the control regions of several flounder species are approximately 1500 bp with a long insertion between CSB-III and tRNA-Phe. Since the unexpected extension of sequences could be due to the duplication of large segments, I searched for repetitive sequence units in the flounders. Winter and yellowtail flounders have 6-8 units of repetitive sequences in the tRNA-Phe flanking region. The sequences of the repetitive units are as follows:

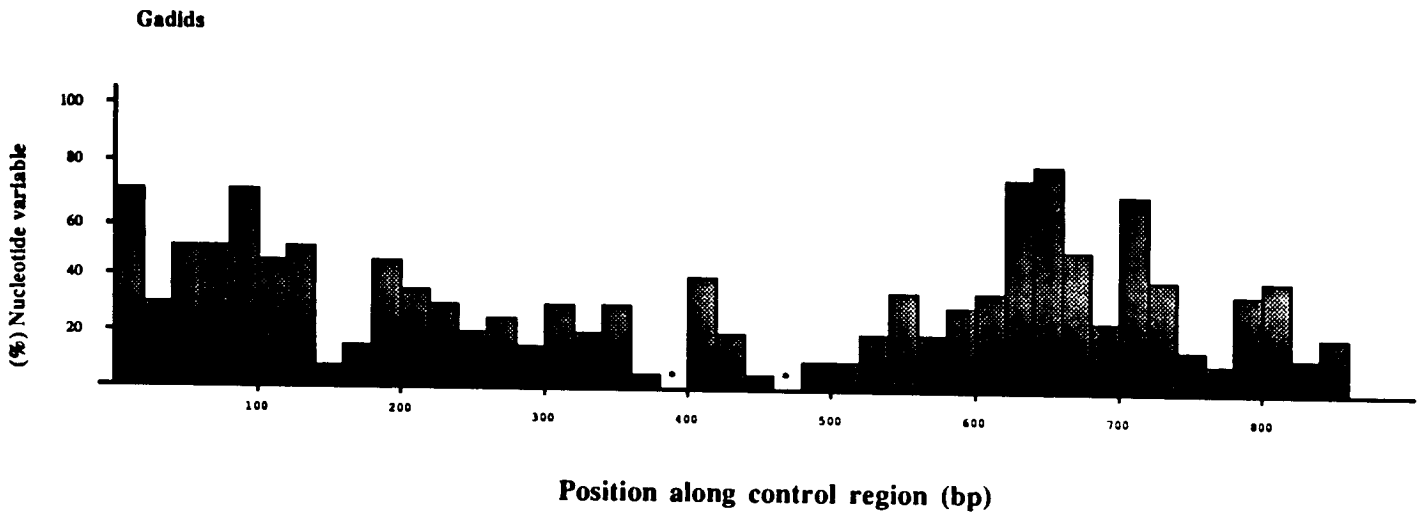


Figure 4-4. Distribution of variable sites along the entire D-loop sequences of four gadids. Each bar represents 20 base segment and asterisks (*) denote the most conserved regions. Sequences in the central region are generally the most conserved in these fishes.

```
5'-A---TTCCAACCCTAAAATAACGGTGACCTACC-  (yellowtail)
5'-ACTTTTTCAACCCTAAAATAGCGGTAGCCCACC- (winter flounder)

  -CCAGACTTTCCAATTATTGTAATAATGTTAAT-3' (yellowtail)
  -CCAAGCTTACCGATTATTGTAATAATGTT----3' (winter flounder)
```

The winter flounder repeat disrupts a highly conserved sequence block (CSB-III), a result I checked by direct sequencing of the region. In both species, the repeat sequence is related to weakly conserved sequences adjacent to CSB-III. The exact numbers of the repetitive sequence units is not known, and is probably variable among individuals. The size of our amplified products suggests that there are eight or nine copies of the repeat in each species. No sequence variation was observed among repeat copies within species, implying a relatively high rate of homogenization of the repeats.

The greater length of the American plaice control region is due to an independent insertion of 393 nucleotides. The insert contains two copies of a 61bp sequence with only 9 mismatches between the copies. No other regions of significant similarity are present. All three species share a conserved sequence block adjacent to tRNA-Phe which probably corresponds to the promoters.

Labroidei

The first half of the labroid control regions were amplified with primer pair (A,E or K,E). These sequences share many similarities with pleuronectids, consistent

with their placement together in the Percomorpha. A segment of approximately 40 bp near the beginning of the sequence is shared among all of these families. Potential secondary structures in this region include stem-loop structures, suggesting that these are termination-associated sequences (Doda et al., 1981). After the conserved region, a segment of about 200 bp is highly variable among Caribbean reef fishes, implying this region might be the best for the population studies of these species.

The complete sequence of the African cichlid control region is 888 bp long (Cichlid-A). Another cichlid species is from South America (Cichlid-B). Indels between two sequences are mostly single base pair events. The most variable portion of this sequence also lies in the first 300 bp before the central conserved region.

Sea Lamprey

Since the control region of sea lamprey mtDNA carries tandem repeats of 117 bp in the first half, the remaining part is just 374 bp, which is by far smaller than any other vertebrates sequenced so far. The tandem repeats are well formed into the secondary sturucture as shown in the cod repeats. The sequence of the lamprey control region shows very low similarity to other teleosts and therefore the sequence alignment is highly ambiguous. Only two conserved sequence blocks are identifiable. The conserved sequence domains are CSB-II, and III. The sequence are compared with loach D-loop sequence.

```
CSB-II :  Loach    5'- CGACAAACCCCCTTACCCCCTT -3'
          Lamprey  5'- CGACAA-CCCCCTTACCCCCTT -3'
```

```
CSB-III:  Loach     5'- TCAAACCCCGAAAGG -3'
          Lamprey   5'- TCAA-CCCCCTTAGG -3'
```

Of the two CSBs, CSB-II is the most conserved with only one indel.

Potential Secondary Structures of CSBs

The control regions of mammals contain a number of CSBs (Southern et al., 1988; Hoelzel et al., 1991). Although the function of most of these is not understood (Kocher et al,, 1991), their conservation across vertebrates suggests an important role in mitochondrial metabolism. The complete control region sequences I have obtained from four gadids, alewife, plaice, carp, and salmon, provide another opportunity to search for conserved secondary structures. The sequences were analyzed with FOLD program to observe the homology of the structures in the most conserved sequence blocks. Although I did find potential secondary structures with long stems I did not observe compensatory base substitutions among species which might confirm the reality of the structures.

Phylogenetic Relationships

A UPGMA clustering derived from the PILEUP alignment of sequence characters from the central conserved region is shown in Figure 4-5. Species in the same family are consistently clustered (e.g. Gadidae, Cyprinidae, Pleuronectidae), suggesting that the control region does carry phylogenetic information at least to the

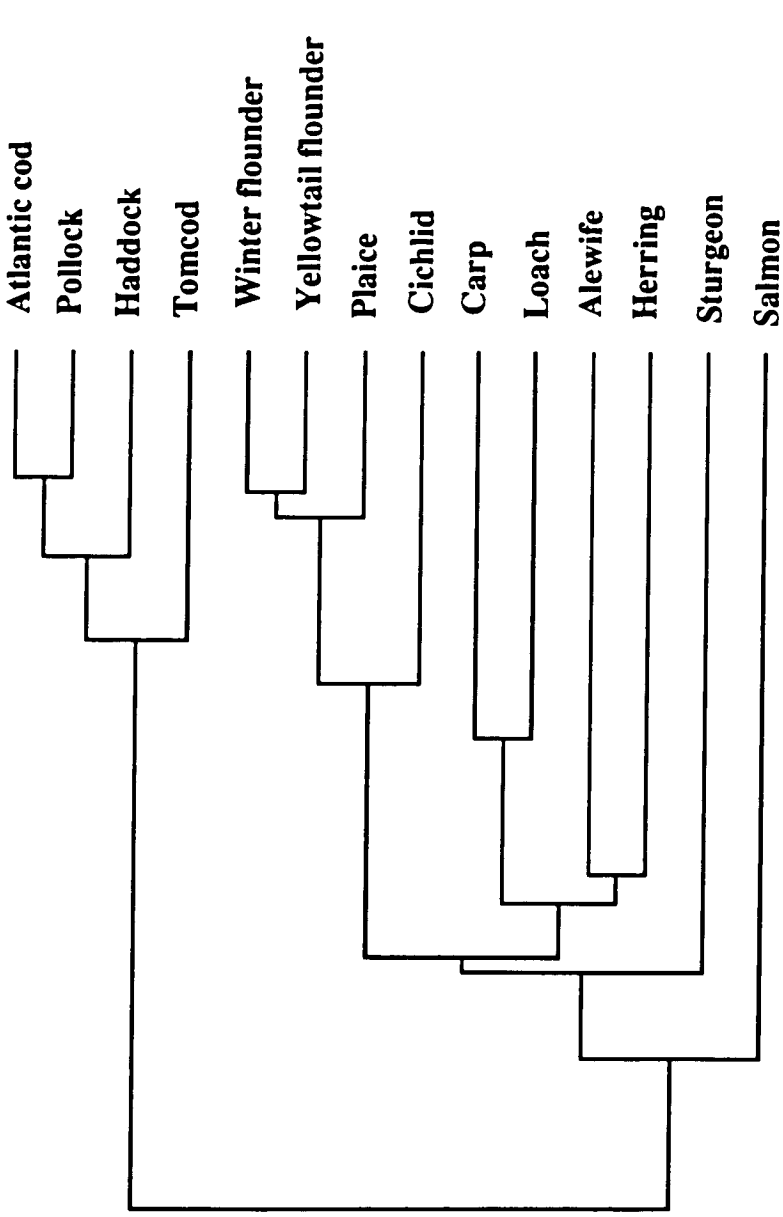


Figure 4-5. UPGMA clustering produced by PILEUP for the central conserved region sequences of 14 fish species

family level, despite its fast evolutionary rate. The monophyletic clustering of cichlids with the pleuronectids is consistent with the definition of Acanthopterygii (Lauder and Liem, 1983). Bootstrapping of parsimony analyses based on the same characters reaffirms the familial clustering, but provides little support for higher-level relationships.

DISCUSSION

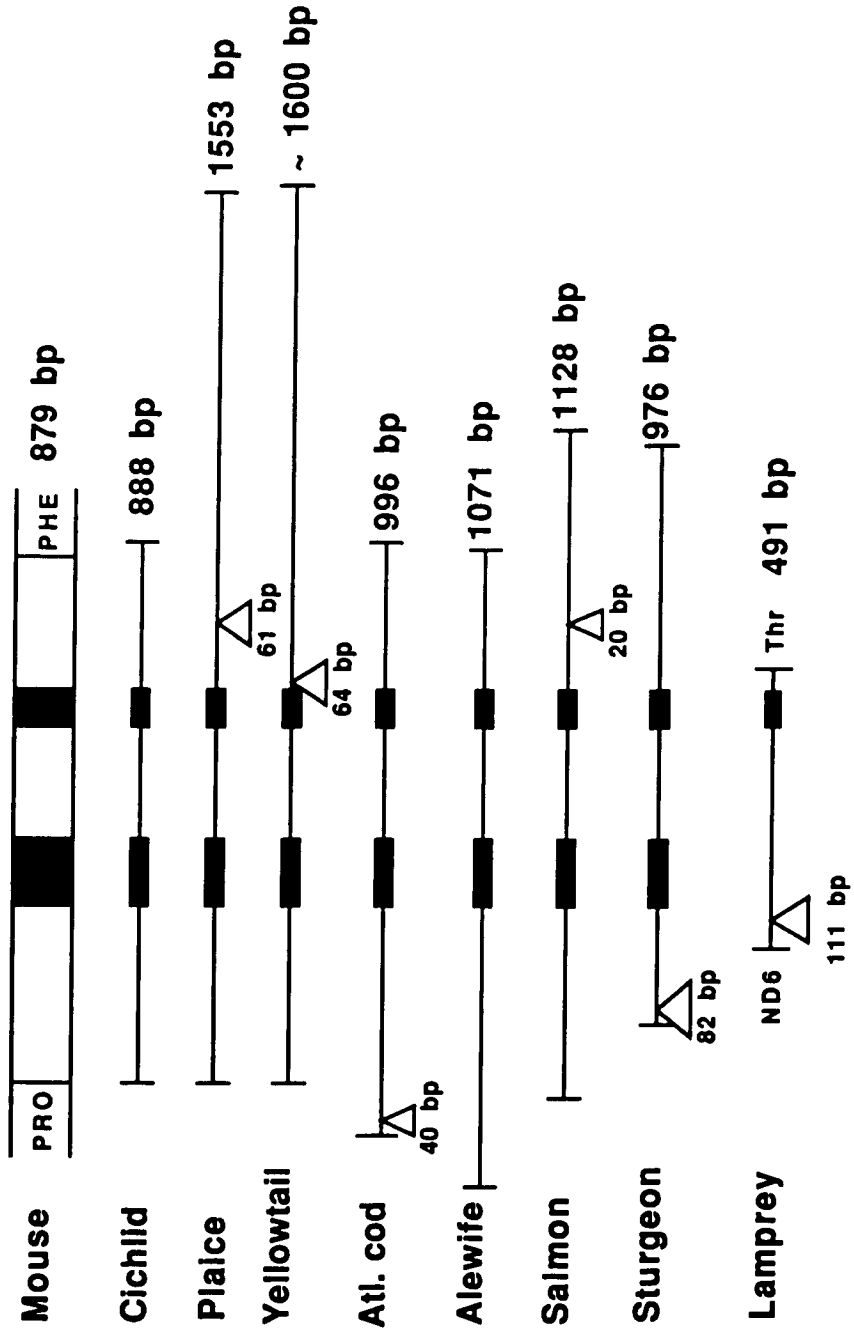
The primary objective of this study was to identify the most suitable segment for detecting intraspecific haplotype variation. I expect that the region showing the highest level of interspecific difference will also be the most variable within species. Alignment of the sequences with respect to the central conserved region shows that the first half of the control region exhibits considerable size variation among species (Fig. 4-6) due to numerous small indels. This region also shows high levels of nucleotide substitution (Fig. 4-4), and is probably the most useful for studies of intraspecific variation.

The second half of the yellowtail and winter flounder control regions show an extraordinary exaggeration in length, due to the presence of long repetitive sequences. Individual fish may frequently be heteroplasmic for variable numbers of these repeats, and so these segments are difficult to sequence. This feature makes the region less attractive for population studies. The plaice, however, contains a large unique sequence insert adjacent to the promoter, which may be particularly useful.

Functional Constraints

The alignments (Fig. 4-3) allowed us to identify conserved sequences which are presumably under function constraint. A central conserved region, roughly

Figure 4-6. Overview of the aligned control region sequences from seven fish species, lamprey and a mammal. Length variation among species is not confined to the first half of the sequence. Repetitive sequences are found most oftern near the ends of the D-loop strand.



homologous to mammalian conserved sequence blocks (CSB's) C-F, is the most obvious feature. In this region, CSB-D is the most universally conserved segment among fish families but not in lamprey, suggesting that the central conserved domain including CSB-D is a recently arisen block. The CSB-II and III which are located outside the central conserved domain, are the most conserved in lamprey. This features support the hypothesis that the replication of the first strand initiates around CSB-II and the CSB-II is an important site for the primer. Although a number of different approaches have been tried (Mignotte et al., 1987; Saccone et al., 1987), the function of this central conserved region is not understood. As mentioned in previous D-loop studies in mammals (Hoelzel et al., 1991) and in *Xenopus* (Wong et al, 1983), potential secondary structures of the control region sequences are not conserved among vertebrates. I conclude that secondary structure of RNA does not play a major role in the function of these sequences, and suggest that these regions may instead represent sites of protein binding to the DNA helix. Furthermore, it is obvious that the CSB-D, II, and III may play the most important roles in the vertebrate mitochondrial control region and the rest of CSBs are probably recent inventions.

Gadid control regions show substantial differences in the central conserved region, including the lack of a GTGGG-box and numerous deviations from the teleost consensus. These differences imply rapid evolution of structural interactions of this regulatory sequence which may be useful in experimental dissection of structure/function relationships of this region.

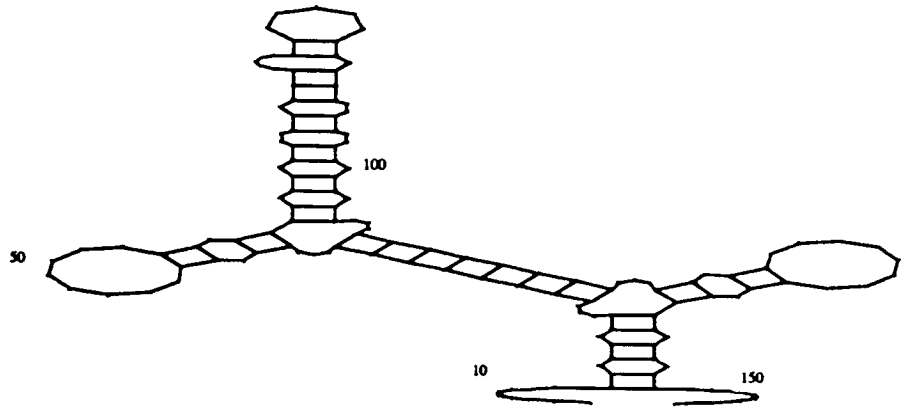
Origin of Repeated Sequences

Five species contained distinctive repetitive sequences in the control region. The Atlantic cod I sequenced had four identical repeats in the beginning of the control region. The yellowtail and winter flounder had approximately 8 identical copies of a 64 bp sequence in the second half of the control region. The plaice had a long (393 bp) insert containing two, nonidentical repeats of a 61 bp sequence. Sturgeon has been previously reported to contain a variable number of an 82 bp repeat adjacent to tRNA-pro (Buroker et al., 1990)

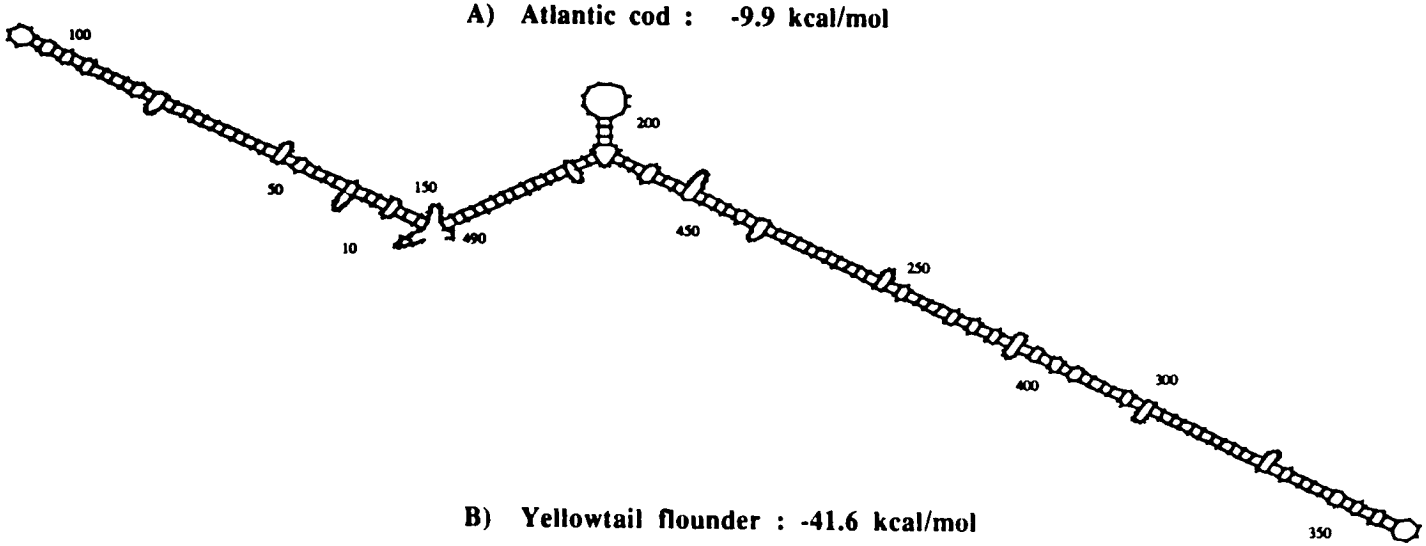
The sturgeon and cod repeats are located in a region in which the D-loop strand is terminated. Buroker et al. (1990) proposed that such repeats arise and are maintained by stable secondary structures which induce misalignment prior to elongation of the D-loop strand for replication. In support of this hypothesis, the Atlantic cod repeat can be folded into a stable secondary structure (Fig. 4-7).

The flounder repeats are located in the region in which replication is initiated. Normally, transcripts originating from the promoters are responsible for priming replication (Chang and Clayton, 1986). In mammals, the transition to DNA synthesis occurs in the region surrounding CSB-II. The position of the flounder repeats immediately upstream of this CSB suggests that they may be maintained by a similar mechanism as replication of the heavy strand nears completion. In support of this view, the yellowtail and winter flounder repeats do exhibit the potential to form stable secondary structures (Fig. 4-7). Although there is clear evidence for a repetitive

Figure 4-7. Secondary structures of the repetitive sequences found in fish control regions. a) sequence from the first half of the Atlantic cod control region b) sequence from the second half of the yellowtail flounder control region.



A) Atlantic cod : -9.9 kcal/mol



B) Yellowtail flounder : -41.6 kcal/mol

structure in the plaice control region, it appears that these repeats are no longer homogenized. Sequence divergence now makes it difficult to reconstruct potential secondary structures in this species.

Phylogenetic Insight

As expected, control region sequences have little utility for analysis of deep phylogenetic divergences. The main difficulty is the identification of homologous sites. Because this non-coding region lacks the strict triplet structure of protein-coding genes, many more indels are tolerated by selection. It becomes difficult to align sequences after only a few million years of divergence.

Only in the central conserved region is there enough selective constraint on a continguous group of nucleotides to allow a reasonable hypothesis of homology to be made. A UPGMA clustering of sequences, as derived by the PILEUP algorithm, accurately reconstructs relationships at the family level (Fig. 4-5). PILEUP clustering of the first half of the control region demonstrates that the Central American and African cichlids are as different from each other as cunner and redfish are from each other. This underscores the great age of the family Cichlidae. While I place little confidence in the higher-level relationships developed from the present data, a careful analysis of more appropriate regions of the mtDNA molecule may be useful for studying relationships among teleost groups.

Acknowledgements

I thank Phil Levin and Mike Amstrong for the donations of some samples.

CHAPTER V

EVOLUTIONARY RATES OF MITOCHONDRIAL DNA AMONG MAJOR VERTEBRATE LINEAGES INFERRED BY ANALYSIS OF AMINO ACID SEQUENCES

ABSTRACT

Using amino acid sequences of eleven mitochondrial protein genes from ten vertebrates, the phylogeny and estimated rates of evolution among vertebrate lineages were obtained. The tree topology for the eutherian orders is identical to that of a previous study. The estimated rate of sequence divergence in the warm-blooded vertebrates is generally faster than in the cold-blooded vertebrates. Among cold-blooded vertebrates, the lamprey lineage appears to evolve the fastest. Based on the comparisons of two recently diverged groups, mouse and rat versus carp and loach, the evolutionary rate in the warm-blooded vertebrates is estimated two to five times faster than in cold-blooded counterparts. However, the uncertainty of fossil records can contribute large errors to the estimation of evolutionary rates. Moreover, the amino acid sequence substitutions quickly saturate around the time point of 100 million years after divergence, which is about the same rate as observed in nucleotide transversion rate of the third positions of mitochondrial protein coding genes.

Conclusions about the relative rates of amino acid substitution in various groups should wait for methods which account for the pattern of sequence saturation, as well as additional sequence data which may reveal the complete record of substitution in these genomes.

INTRODUCTION

Since the development of the evolutionary clock concept in the 1960's, molecular data have been widely used to infer phylogenetic topology as well as divergence time between taxa. Animal mtDNA has provided unique perspectives for evolutionary studies because of its high rate of evolution (Wilson et al., 1985). The divergence time estimation between taxa often relies heavily on an assumption of homogeneity in evolutionary rate among different taxonomic groups under study. One challenge to using molecular data appropriately is the suggestion that evolutionary rate of the mitochondrial DNA varies among major vertebrate taxa. For instance, it has been suggested that the rodent lineage evolves faster than other mammalian lineages (Li et al., 1986) and that the rate is much slower in sharks than in primates (Martin et al., 1992). Heterogeneity in evolutionary rate has also been suggested from comparisons of amino acid sequences of mitochondrially encoded proteins from several vertebrates. Warm-blooded vertebrates appear to accumulate substitutions six times faster than cold-blooded vertebrates (Adachi et al., 1993). The higher rate of evolution in warm-blooded animals has been attributed to relaxed selective constraints of proteins in a thermally stable environment or higher levels of mutation correlated with a higher metabolic rate.

The use of protein sequence rather than nucleotide sequence, particularly for

distantly related taxa, has been encouraged, because of the following reasons. Saturation of the amino acid changes is not as evident as that of nucleotide substitutions between distantly related taxa. Moreover the nucleotide composition of the genome has a minimal effect on the amino acid sequence because mutations are filtered at the codon level (Reeves, 1992; Cao, 1993). It has been reported that the maximum likelihood (ML) method is generally robust with respect to heterogeneity of evolutionary rate among taxa (Hasegawa et al., 1993). While I suspect that this is true for conclusions about topology, I doubt whether it is true for estimates of branch length.

Since the maximum likelihood method for amino acid sequences totally depends on the evolutionary model, three different models (JTT, Dayhoff, and Poisson) are frequently employed to infer relationship between taxa (Reeves, 1992; Hasegawa et al., 1993). The JTT model is an empirically transitional matrix derived from a dataset 40 times as large as the original Dayhoff model (Dayhoff et al., 1978).

In the construction of phylogenies using molecular data, heterogeneity in evolutionary rate has some potentially important implications. Models of the substitution process should take these discrepancies into account when inferring topologies. The Poisson model is the simplest and can be explained by two parameters as follows: The probability of the current amino acid i being replaced by another amino acid j is based on the rate of amino acid substitutions, u in an infinitesimally short period of time (dt), $P_{ij} = udt$. The recent PROTML version

(2.1.2) has another option called 'F' that stands for the Proportional model. Because of the discrepancy between the averaged transitional matrix and the actual amino acid substitutions, the incorporation of the proportional model into the three basic models reduced the discrepancy (Cao et al., 1993). The model minimizing the Akaike's information criterion (AIC) is considered to be the most appropriate model (Adachi et al., 1993).

Recently, mitochondrial genomes from two bony fishes (Tzeng et al., 1992; Chang et al., 1994) and one agnathan (chapter III) have been sequenced, providing an opportunity to test rates of evolution among cold-blooded vertebrates. Although the tempo and mode of evolution for mitochondrial DNA sequence have been relatively well documented, little is known about the evolutionary pattern of mitochondrially encoded protein sequence. The main goal of this study is to estimate the evolutionary rates among different vertebrate lineages using protein sequences. In this chapter, I will compare the protein evolutionary models, and the rates of evolution among the major vertebrate lineages will be compared and discussed.

MATERIALS AND METHODS

Ten vertebrate species were examined for phylogeny and evolutionary rate tests using sea urchin as an outgroup. The amino acid sequences encoded by mtDNA from human (Anderson et al., 1981), cow (Anderson et al., 1982), mouse (Bibb et al., 1981), rat (Gadaleta et al., 1989), chicken (Desjardins and Morais, 1990), frog (Roe et al., 1985), carp (Chang et al., 1994), loach (Tzeng et al., 1992), and sea urchin (Jacobs et al., 1988) were obtained from the Genbank data base and aligned with ESEE and PILEUP in the GCG package. Of 13 protein genes, all genes encoded on the H-strand except ATP8 gene were used (Fig. 5-1). The ATP8 gene is the shortest of the mtDNA protein-coding genes and the most variable, making the alignment too ambiguous.

As mentioned in Adachi et al. (1993), there are also several exclusions in the data used for this study. The initiation codons, stop codons, overlapping regions, and unaligned sites at the ends of genes were not included. However I did not exclude sequences from the middle of genes. Phylogenetic trees, distances from three different models (JTT-F, Dayhoff-F, and Poisson-F) and the number of base substitutions between taxa were obtained by PROTOML (ver. 2.1.2) or MEGA (ver. 1.0).

The evolutionary rates across vertebrate lineages were measured by both branch length and the difference in the number of amino acid substitutions versus the

divergence time estimated from paleontological records (Carroll, 1988; Colbert et al., 1991).

Figure 5-1. The aligned amino acid sequences used in this study. OTUs 1 to 11 respectively represent human, cow, mouse, rat, marsupial, chicken, frog, carp, loach, lamprey, and sea urchin. All protein sequences encoded on the first strand except ATP8 were used. Dashes (-) denote gaps introduced during alignment and dots (.) represent identical amino acids compared to the human sequence. Each gene is indicated above the sequences and some highly variable sequences were excluded.

OTU	Sequence
	| ND1 →
1	ANLLLLIVPI LIAMAFLMLT ERKILGYMQL RKGPNVVGPY GLLQPFADAM KLFTKEPLKP
2	I.I.M..I.. .L.V...T.V ...V..... ....I...I ...I....R.
3	I.I.T.L... ....T.V ....I.... .I..... ...M...MR.
4	I.I.T.LI.. ....GL.T.V ....NE... .K..... ...M...MR.
5	I...MY.I.. .L.V...T.V ...V.....F ......I... .I.......L ...I....R.
6	IMT.SY.L.. ...V...T.V .....S...A .....I...F .....V..GV ...I...IR.
7	I.P..YMI.. .L.V...T.I ...V.....H .....I...T ..I..I..GV ...I...VR.
8	I.P.AY...V .L.V...T.I ...V..... ....I..GV ...I...VR.
9	I.P.AY...V .L.V...T.L ...V..... .....I.... .....I..GV ...I...IR.
10	TST.I.VLMV .L.V...TMV ...T..... ....FM .....I..GV ...L...VW.
11	LE.ISFLI.. .LSV...T.V ...V.....F .N.......F ....G. .V.I..E...

1	ATSTITLYIT	APTLALTIAL	LLWTPLPMPN	PLVNLNLGLL	FILATSSLAV	YSILWSGWAS
2	...SASMF.L	..IM..GL..	TM.I.....Y	..I.M...V.	.M..M.....	....
3	L.TSMS.F.I	....S..L..	S..V.....H	..I.....I.	........S.	....
4	L.TSMS.F.I	....S..L..	S..I.....H	..I.....MP	........S.	....
5	M..S.SMFTI	.......L.F	TI........	A.LD......	....L.G.S.	....
6	S..SPF.F.I	T.I...LL..	TI.V...L.F	..AD......	.L..M...T.	..L.......
7	S..SQ.MFLI	...M..AL.M	SI.A.....F	S.AD....I.	....L.....	.T..G...S.
8	S..SPF.FLA	..V....L.M	T..A.M...H	.VTD....I.	....L.....	....G....L
9	S.ASP..FLV	T.M....L.M	T..A.M...H	.IID....V.	....L.....	....G.....
10	LAASPI.F.V	..IM...L..	S..MLI...Q	SISTI.IT..	V.M.I...S.	.A..G.....
11	VN.SPY.FFF	S.L.F.AL..	...NFM.VHT	.TLD.Q.S..	LV.GL...S.	.A..G.....

1	NSNYALIGAL	RAVAQTISYE	VTLAIILLST	LLMSGSFNLS	TLITTQEHLW	LLLPSWPLAM
2	..K.......	....	V	.......T..	.......QM.	.I..A.....
3	..K.S.F...	....	..M......V	...N..YS.Q	........M.	....A..M..
4	..K.S.F...	....	..M.LY...V	.......S.Q	M.......I.	..I.A..M..
5	..K.......	....	I	M.IN...T.K	NMLI...NM.	.IMMT...T.
6	..K.......	....	....	IML..NYT..	..AI...PIY	.IFSA.....
7	..K.......	....	...GL...CM	IMLA.G.TYT	..M....QM.	.II.G..M.A
8	..K.......	....	.S.GL....V	IIF..GYT.Q	.FNI...SI.	..I.A....A
9	..K.......	....	.S.GL..R.I	IIFW.GYTVQ	.FN....A..	....AC...A
10	..K.......	....	.S.GL...CL	VILT...S.Q	AF.Y....T.	F..S.....A
11	..K.S.L..I	....	IS..L....L	IIF.S....T	YIMN...FS.	FS.SCL..FY

1	MWFISTLAET	NRTPFDLAEG	ESELVSGFNI	EYAAGPFALF	FMAEYTNIIM	MNTLTTTIFL
2	....	..A....T..	V	....	.....A....	..IF.AIL..
3	....	..A....T..	V	....	L	..A...I...
4	..Y.......	..A....T..	V	....	L	..A..SIV..
5	..Y.......	..A....T..	V	........M.	.L...A..MV	..AI.A.L..
6	..Y.......	..A....T..	V	........M.	.L...A..ML	......VL..
7	..Y.......	..A....T..	V	...G......	SL...A..L.	....SYL.LF
8	..Y.......	..A....T..	V	...G......	.L...A..LL	....SAVL..
9	..Y.......	..A....T..	....AP...V	...G.....L	.L...A..LL	....SAIL..
10	...V......	.......T..	V	...G......	.L...S..LF	.....AIM..
11	I..V......	..A....T..	...I...Y.V	...G...V..	.I...A...L	..YFSVVL..

1	GTTYDALSPE	LYTTYFVTKT	LLLTSLFLWI	RTAYPRFRYD	QLMHLLWKNF	LPLTLALLMW
2	..SHNPHM..	...IN.TI.S	....MS....	.AS.......	....	.......C..
3	.PL.YINL..	..S.N.MMEA	...S.T....	.AS.......	....	.......C..
4	.PL.HINY..	..S.S.M.E.	...STT....	.AS.......	....	......FC..
5	.SPLSSNI.Y	INSMT.MM.M	.I..TT....	.AS.......	...Y......	..I....CL.
6	NPSFLN.P..	.FPIALA...	...S.S....	.AS.......	....	.......CL.
7	LGSSFMNQ..	.T.ISLMI.S	SI.SMI...V	.AS.......	.....V....	..I...MTL.
8	.ASHIPSV..	.T.INLM..A	A..SIM...V	.AS.......	.....V....	......FVL.
9	.ASHMPAI..	.T.INLM..A	A..SVV...V	.AS.......	.....V....	.......VL.
10	.PLGSNNLNI	.PIINIMM.A	TP.II.....	.AS.......	.....M....	...N...FTL
11	.GPSPLNNLF	PISIII.GIK	TTFLFSV..V	.A........	...F.T..SY	...SIGA.CA

(Figure 5-1 continued)

```
                ← ND1 || ND2 →
  1   YVSMPITISS IPPQNPLAQP VIYSTIFAGT LITALSSHWF FTWVGLEMNM LAFIPVLTKK
  2   H..L..LT.G ......IIFI I.LL..ML.. I.VMI....L LV.I.F.... ..I..IMM.N
  3   HI.L..FTAG V..Y..ITLA I..F...L.P V..MS.TNLM LM.....FSL ..I..M.IN.
  4   .I.L..FLAG ...Y..ITLT I..L.T.K.R ...T..TNLP PM......SL ..I..L.AN.
  5   .I.I..SL.. L...S.YVLT IMSFSLLL.. TM.LI.N..L TA.M...I.T ..I..LM..P
  6   HT....SYAG L..I..H.KL ICTVSLIM.. S..IS.N..I LA.T...I.T ..I..LIS.S
  7   HI.L..SMLG L.S...ITFS .VLTSLASEQ FLAVS....L LA.M...I.T ..I..LM.QH
  8   HTAL..ALAG L.....YRRA TLLCSVGL.. TL.FA....L LA.M...I.T ..IT.LMAQH
  9   HTAL..AFAG L.....YVLM ILI.SLGL.. TL.FA....L LA.M...I.T ..ILRLMAQQ
 10   QL.LAVSFGG AGVPS..I.S TLLM.LGL.. .V.FS.TS.I LA.I...I.T I.I..LMA.T
 11   ILALVALLGI SL.LRQIVST FLFV.VVS.. I.VVS.EN.. II.....LST ..LV.I.CSG

  1   MNPRSTEAAI KYFLTQATAS MILLMAILFN NMLSGQWTMT NTTNQYSSLM IMMAMAMKLG
  2   H...A...ST ......S... .L.M..VII. L.F.....VM KLF.PMA.ML MT..L.....
  3   K........T ...V...... ..I.L..VL. YKQL.T.MFQ QQ..GLILN. TL..LS....
  4   KS.......T .......... ..I.LV.IL. YKQ..M..LQ QQ..NMLLN. MLISL.....
  5   HH...M.S.. ...MI..... ..I.FSAI.. ASTTN..MTG QIS.TSA.F. MTI.L.....
  6   HH..AI..T. ......S... ALI.FSSMT. AWST...DI. QLNHPT.C.. LT..I.I...
  7   KH..AI..ST .......A.. AL..FSS.N. AW.T.E.SIL DL..PL.CAT MTI.IC....
  8   HH..AV..TT .........A AMI.F.STT. AWMT.E.SIN DLS.PIA.T. F.A.L.L.I.
  9   HH..AV..TT .........A AMI.F.ATT. AWAT.E.DIN .L.HPLA.SL T...L.L.V.
 10   HH...I..TT ...IA.SAG. AT..ITACLT AWY..N.AIS PSNDPIILNA MTL.LML...
 11   FS..NV..DN ....V..SSA AL..NGA.GQ AW.T.S.SIL DPV.EVT.IC LSI.L.F.I.

  1   MAPFHFWVPE VTQGTPLTSG LLLLTWQKLA PISIMYQISP SLNVSLLLTL SILSIMAGSW
  2   .......... ....I..S.. .I........ .M.VL...F. .I.LN.I... .V...LI.G.
  3   L......L.. ....I..HM. .I......I. .L..LI..Y. L..STII.M. A.T..FM.A.
  4   L....Y.L.. ....I..HI. .I......I. .L..L..FYQ L..PTITTI. A.S.AFV.A.
  5   L.......... ....I..L.. M.......I. ....F..... ...M...MI. ..T.TLL.G.
  6   LV.....F.. .L..SS.ITA ...S.LM..P ..TLLLLT.Q ...TT..TL. A.S.TLI.G.
  7   L......L.. .L..LS..T. .I.S...... .MA.L...A. M..TP..... GLT.TLI.G.
  8   L..M...M.. .L..LD.LT. .I.S...... .LALII.TAQ TIDPL..TL. G.S.TLV.G.
  9   L..M...M.. .L..LD..M. .I.S...... .FALIT.MA. NT.PM..T.. GL..TLI.G.
 10   ...M...L.. .MV.LDFIT. MI.A...... ..TLLI..AQ DQ.NMFI.IP AL..VFV.G.
 11   L..V...F.D .L..L.FFQ. .IIA....I. .LIL.FYF.Q LGFSY..I.P .LI.VLI.G.

  1   GGLNQTQLRK ILAYSSITHM GWMMAVLPYN PNMTILNLTI YIILTTTAFL LLNLNSSTTT
  2   .......... .M.....A.. ...T...... .T..L...I. ...M.S.M.T MFMA..T...
  3   .......M.. .M.....A.. ...L.I.... .SL.L...M. .....APM.M A.M..N.M.I
  4   .......T.. TM..P..A.. .G.T.I.... .TL.L..... N.L.KAPM.I T.MT.PA..I
  5   .......... .......A.. ...AIIIMIY .SL.....IL .LAS.I.M.M V..QS...KI
  6   M......T.. ...F...S.L ...IMIIS.. .QL...TFIL .T.M.S.V.. S.AQIKVLKL
  7   .......... ...F...A.L ...ISI..FS .QLM...... .L.M.S.M.. V.KTI...KI
  8   .......... .......A.. ...II.IQ.A .QL.LIA.GT ...M.SA... T.KMSLT.KI
  9   .......... .......A.. ...III.Q.T .QL.LIA.GL ..FM.SA... S.KMA.A.KM
 10   .......T.. .......A.. ..ITSMA.F. .TI.W.TTL. .CLI.SAT.I N.HILKANKI
 11   .......V.. ...F...GN. ..LVITSA.S F.AA.IM.V. .L.IN.SL.. .FDHLKVS.L

  1   LLLSRTWNKL TWLTPLIPST LLSLGGLPPL TGFLPKWAII EEFTKNNSLI IPTIMATITL
  2   .S..H....T PIM.V..LA. ...M...... S..M...M.. Q.M.....I. L..F..ITA.
  3   NSI.LL...T PAMLTM.SLM .......... .......I.. T.LM...C.. MA.L..MMA.
  4   NT..PM...T PMILTMASII .......... .......... S.LL...CST LS.L..IMA.
  5   NS..IL...S APNMII.TL. .......... ...M...L.L Q.LINF.NIP LAMML.LS..
  6   ST.LIS.T.T PM.NATVML. ....A..... ...M...L.. Q.L..QEMTP MA..ITMLS.
  7   SS.ATS.S.T PST.A.SLL. .......... S..V...F.. Q.L.SQ.TT. LA.TL.LSA.
  8   ST.AT..S.S PI..STTALV .......... ...M...L.L Q.L..QDLP. .A.A..LTA.
  9   NT.TAA.S.S PV.VSTTAL. .......... ...M...L.L Q.L..QDLPA TA....LAA.
 10   TA.TMNKHNQ ISQML.LLL- .......... ...IN.LLAS I.LANQ.LI. YLFM.MMGS.
 11   GH.NTISQLS PISVA.VLLV M......... ...IL.FTSL YFLVA..FI. LSS..IIGN.
```

(Figure 5-1 continued)

```
                                  ← ND2 || COI →
 1    LNLYFYLRLI YSTSITLLPM SNNFADRWLF STNHKDIGTL YLLFGAWAGV LGTALSLLIR
 2    ......M..T ...TL.MF.S T...IN.... .......... .........M V.........
 3    ...F..I... ....L.MF.T N...IN.... .......... .........M V.....I...
 4    .S.F..T... ..M.L.TF.T N...VN.... .......... .........M V.....I...
 5    ...F..M.I. ..STL.MF.S I...IN.... .......... .........M V.........
 6    .S.F.....A .HST...P.N .S..IN.... .......... ..I..T...M A.........
 7    .S.F.....T .IVTL.SS.N TS.AIT.... .......... ..V......L V.........
 8    IS.......L .AMTL.IS.N MI.AIT..F. .......... ..V......M V.........
 9    .S.......S .AMAL.IY.N TT.AIT..F. .......... ..V......M V.........
10    .S.F..T.MC .LSI.LSP.C .TTTHI.... .......... ..I......M V.....I...
11    QDYF....IS FN..LF.F.Q HIIQLS.... .......... ..I......M V...M.VI..

 1    AELGQPGNLL GNDHIYNVIV TAHAFVMIFF MVMPIMIGGF GNWLVPLMIG APDMAFPRMN
 2    .......T.. .D.Q....V. .......... .......... .......... ..........
 3    .......A.. .D.Q..... .......... ....M..... .......... ..........
 4    .......A.. .D.Q..... .......... ....M..... .......... ..........
 5    .......T.I .D.Q..... .....I.... .......... .......... ..........
 6    .......T.. .D.Q..... .......... .......... .......... ..........
 7    ...S...T.. .D.Q..... .....I.... .......... .......... ..........
 8    ...S...S.. SD.Q..... .......... .....L.... .......... ..........
 9    ...N...A.. .D.Q..... .......... .....L.... .......... ..H.......
10    ...S...T.. .D.Q.F.... .......... ..I....... ........L. ..........
11    ...A...S.. ND.Q....V. ....L..... .......... ....I..... ..........

 1    NMSFWLLPPS LLLLLASAMV EAGAGTGWTV YPPLAGNYSH PGASVDLTIF LHLAGVSSIL
 2    .......... F.......S.. .......... .......LA. A......... ..........
 3    .......... F.......S.. .......... .......PV. A......... ..........
 4    .......... F.......S.. .......... .......LA. A.V....... ..........
 5    .......... F.......STI .......... .......LA. A......A.. .....I....
 6    .......... F.......ST. .......... .......LA. A......A.. HY........
 7    .......... F.......SG. .......... .......LA. A......... .....I....
 8    .......... F.......SG. ........S. .......LA. A......... ..........
 9    .......... F.......SG. .......... .......LA. A......... ..........
10    .......... ........G. .......... .......LA. T......... ..........
11    ......I... FI......G. .N.......I ....SS.IT. A.S....A.. .....A....

 1    GAINFITTII NMKPPAMTQY QTPLFVWSVL ITAVLLLLSL PVLAAGITML LTDRNLNTTF
 2    .......... .......S.. .........M .......... .......... ..........
 3    .......... .......... .......... .......... .......... ..........
 4    .......... .......... .......... .......... .......... ..........
 5    .......... .......S.. .........M .......... .......... ..........
 6    .......... ......LS.. .......... ...I...... .......... ..........
 7    ........T. .......S.. .......... .......... .......... ..........
 8    ........T. ......IS.. .......... V......... .......... ..........
 9    ........T. ......LS.. .......A.. .......... .......... ..........
10    ..V......F .....T.... .......... .......... .....A.... .........S.
11    .L........ ..RT.G.SLD RL.......F V..F...... ....GA.... .....I....

 1    FDPAGGGDPI LYQHLFWFFG HPEVYILILP GFGMISHIVT YYSGKKEPFG YMGMVWAMMS
 2    .......... .......... .......... .......... .......... ..........
 3    .......... .......... .......... ...I...V.. .......... ..........
 4    .......... .......... .......... ...I...V.. .......... ......T...
 5    .......... .......... .......... .......... .......... ..........
 6    .......... .......... .......... .......V.A ..A....... .........L.
 7    .........V .......... .......... .......... .......... ..........
 8    .......... .......... .......... ...I...V.A .......... .........A
 9    .......... .......... .......... ...I...V.A ..A....... .........A
10    .......... .......... .......... ...I...V.A ..A....... .........A
11    .......... .F.....L.. .......... .......VIA H....R.... .L.L.Y..IA
```

(Figure 5-1 continued)

```
 1    IGFLGFIVWA HHMFTVGMDV DTRAYFTSAT MIIAIPTGVK VFSWLATLHG SNMKWSAAVL
 2    .......... .......... .......... .......... .......... G.I...P.MM
 3    .......... .......L.. ....C..... .......... .......... G.I...P.M.
 4    .......... .......L.. .......... .......... .......... G.I...P.M.
 5    .......... .......L.. .......... .......... .......... G.I...P.M.
 6    .......... ......R... .......... ........I. .......... GTI..DPPM.
 7    ..L....... ......DLN. .......... .......... .......M.. GTI..D.PM.
 8    ..L....... .......... .......... .......... .......... GSI..ETPM.
 9    ..L....... .......... .......... .......... .......... GTI..DTPM.
10    ..L....... .......... .......... .......... .......... GKIV.HTPM.
11    ..V...L... .......... .......A.. ....V...L. ....M.K.Q. ..LQ..LPL.

 1    WALGFIFLFT VGGLTGIVLA NSSLDIVLHD TYYVVAHFHY VLSMGAVFAI MGGFIHWFPL
 2    .......... .......... .......... .......... .......... ....V.....
 3    .......... .........S .......... .......... .......... .A..V.....
 4    .......... .........S .......... .......... .......... .A..V.....
 5    .......... I......... .......... .......... .......... ....V.....
 6    .......... I......... ......A... .......... .......... LA..T.....
 7    .......... .......... ......M... .......... .......... ..........
 8    .......... .........S .......... .......... .......... .AA.V.....
 9    .......... .........S .......... .......... .......... .A..V.....
10    .......... .........S ......I... .......... .......... .A..V.....
11    .T..IV.... L......... ...I.F.... .......... .......... FA..T.....

 1    FSGYTLDQTY AKIHFTIMFI GVNLTFFPQH FLGLSGMPRR YSDYPDAYTT WNILSSVGSF
 2    ......ND.W .....A...V ...M...... .......... .........M ..TI..M...
 3    ...F...D.W ..A..A...V ...M...... .......... .......... ..TV..M...
 4    ......ND.W ..A..A...V ...M...... ....A..... .......... ..TV..M...
 5    .T..M.NDMW .....F...V .......... .......... .........M ..VV..I...
 6    .T.F..HPSW T.A..GV..T .......... ....A..... .........L ..T...I..L
 7    .T....HE.W .....GV..A .......... .....A.... .........L ..TV..I..L
 8    LT....HS.W T....GV... .......... .....A.... ........AL ..TV..I..L
 9    .T.FS.HD.W T....GV... .......... ....A..... .........L ..TV..I..L
10    .T....NE.W ..A..I...A .......... ....A..... .......... ...I..I..T
11    ....S.HPLW G.V..F...V .......... ....A..... .........L ..TI..I..T

                                                    ← COI || COII →
 1    ISLTAVMLMI FMIWEAFASK RKVLMVEEPS MNLEWLYGCP PPYHTFEEPQ VGLQDATSPI
 2    .........V .I........ .E..T.DLTT T.....N... .......... L.F.......
 3    ......LI.. .......... .E.MS.SYA. T.....H... .......... L.........
 4    ......LV.. .......... .E..SISYS. T.....H... .......... L.........
 5    ......I..V .I........ .E..D..LTT T.I....... .......Q.. L.F.....S...
 6    ..M...IMLM .IV....SA. ....QP.LTA T.I..IH... .......... L.F.....S...
 7    ...V..IM.M .I......A. .E.TTY.LT. TM....Q... T....LKTS. L.F...A...
 8    ...V..IMFL .IL.....A. .E..S..LTA T.V...H... .....Y.... L.FK..AM.V
 9    ...V..IIFL .IL....... .Q.MS..LTM T.V...H... .......... L.F...A..V
10    V..I....FM .IL....SA. ..AIATDLLN T.....H... .....Y.... L.....A...
11    ..VV.MLFFL .L.......Q .EGITP.FSH AS...Q.TSF ..S.HTFDE. F.....S..L

 1    MEELITFHDH ALMIIFLICF LVLYALFLTL TTKLTNNISD AQEMETVWTI LPAIILVLIA
 2    ....LH.... T...V...SS ....IIS.M. .....HSTM. ...V..I... .......I...
 3    ....MN.... T...V...SS ....IIS.M. .....HSTM. ...V..I... ...V..IM..
 4    ....TN.... T...V...SS ....IIS.M. .....HSTM. ...V..I... ...V..I...
 5    ....MY.... T...V...SS ....III.M. .....HSTM. ...V..I... ...V..I...
 6    ....VE.... ...VALA..S ....L.T.M. ME..SS.TV. ...V.LI... ....V...L.
 7    ....LH.... T..AV...ST ....IITIMM .......LM. ...I.M.... M...S.IM..
 8    ....LH.... ....VL..ST ....IITAMV S.....Y.L. S..I.I.... ...V......
 9    ....LH.... ....V...SA ....VIIT.V S.....Y.L. S..I.I...V ...L..I...
10    .....H.... T.TVV...SV .IF.LIIVMV ..TFI.HSL. S..V.I...V M...V.IT..
11    ....TY...Y ..IVLT..TI ..F.LVS.LV SSNTNRFFFE G..L..I..V I..L..I...
```

(Figure 5-1 continued)

```
 1    LPSLRILYMT DEVNDPSLTI KSIGHQWYWT YEYTDYGGLI FNSYMLPPLF LEPGDLRLLD
 2    .........M ..I.N....V .TM......S ......ED.S .D...I.TSE .K..E....E
 3    .........M ..I.N.V..V .TM......S ......ED.C .D...I.TND .K..E....E
 4    .........M ..I.N.V..V .TM......S ......ED.C .D...I.TND .K..E....E
 5    .........M ..IYN.Y..V .AM......S ..F...EN.M .D...I.TKD .S..Q....E
 6    ....Q....M ..IDE.D..L .A........ .....FKD.S .D...T.TTD .PL.HF...E
 7    ........LM ......H... .A.......S ....N.ED.S .D...I.TND .T..QF...E
 8    ........LM ..I...H... .AM......S ......EN.G .D...V.TQD .A..QF...E
 9    ........LM ..I...H... .AM......S ......EN.S .D...I.TQD .T..QF...E
10    ........L. ..ISN.H... .AV......S ......HQME .D...I.TNE ....GI....
11    ....QL..LM ......F... .AF......S .....FND.E .D...V.TSD VSF.NP...E

 1    VDNRVVLPIE APIRMMITSQ DVLHSWAVPT LGLKTDAIPG RLNQTTFTAT RPGVYYGQCS
 2    ........M. MT...LVS.E .........S .......... ......LMSS ...L......
 3    ........M. L....L.S.E .........S .......... ....A.V.SN ...LF.....
 4    ........M. L....L.S.E .......I.S .......... ....A.V.SN ...LF.....
 5    ....I...M. L....L.S.E ....A.TM.S ....A..... ....I.L.SS ....F.....
 6    ..H.I.I.M. S...VI..AD .........A ..V....... .....S.IT. ....F.....
 7    ....M.V.M. S.T.LLV.AE .........S ..V....... ..H..S.I.. ....F.....
 8    T.H.M.V.M. S.V.VLVSAE .........S ..V.M..V.. ....AA.I.S ....F.....
 9    T.H.M.V.M. S...ILVSAE .......L.A M.V.M..V.. .....A.I.S ....F.....
10    ..H.I.V.M. S.V..L...E ..I...TI.S ..T.V..V.. ....A..IT. ...LFF....
11    ....L...MQ N...VLVS.A .........S ..T.M..V.. .......F.A .T..F.....

                          ← COII || ATP6 →
 1    EICGANHSFM PIVLELIPLK IFEMGGRTWS LMLVSLIIFI ATTNLLGLLP HSFTPTTQLS
 2    ....S..... ......V... Y..KW.Q..T ...M...L.. GS........ ..........
 3    ....S..... .....MV... Y..NW....T ..I....M.. GS........ .T........
 4    ....S..... .....MV... Y..NW....A ..I....M.. GS........ .T........
 5    ....S..... .....MAS.. Y..KW....T ...M...L.. .S........ Y.........
 6    ........Y. ...V.ST... H..AW.HK.A .L.T...LML LSI....... YT........
 7    .......... ...V.AV..T D..NW.HK.A .L.T..MLLL MSL....... YT........
 8    .......... ...V.AV..E H..NW.HK.A .L.A..M..L I.I.M..... YT........
 9    .......... ...V.AV..S H..NW.HK.A .L.A..MV.L I.I.M..... YT........
10    .......... ..A..AV..S N..NW.HK.A .ICMASMM.. LMI....... YTY.......
11    .......... ...I.SV.FN T..NWTAP.A GLIAGVFVL. LLV.V...F. PYAFQSPTSN

 1    MNLAMAIPLW AGTVIMGRSK IKNALAHFLP QGTPTPLIPM LVIIETISLL IQPMALAVRL
 2    ...G...... ..A..T..N. T.AS...... .......... .........F ..........
 3    ...S...... ..A..T..H. L.SS...... ....IS.... .I.......F ..........
 4    .D.S...... ..A..L..H. L..S....S. ....IS.... .I.......F ..........
 5    ..IG...... ........N. P.MS...... .......... .I.......F ...L......
 6    ..M.L.L... LA.LLT..NQ PSAS.G.L.. E........A .IM...T... .R.L..G...
 7    L.MGL.V... LA....ASKP TNY..G.L.. E........V .I.......F .R.L..G...
 8    L.MGF.V... LA...I..NQ PTI..G.L.. E...I....V .I........ .R.L..G...
 9    L.MGL..... LA...I..NQ PTV..GDL.. E...L....V .I.......F .R.L..G...
10    ..MGL.V... LA..LI.QK. PTE....L.. E...AA.... .I.......F .R.I..G...
11    LTYSLGF... MAIN.L.FYL AF.S.S.LV. ....SA...L M.W...L..F A..I..GL..

 1    TANITAGHLL MHLIGSATLA MSTINPSTLI IFTILILLTI LEIAVALIQA YVFTLLVSLY
 2    .......... I....G.... LMS.STTA.. T......... ..F...M... ..........
 3    .......... .....G...V LMN.S.TAT. T.I..L.... ..F....... ..........
 4    .......... .....G...V LMD.S.TAT. T.I..L...V ..F....... ..........
 5    .......... I......... L.S.STVST. T.S..F...L ......M... ..........
 6    ...L...... IQ..ST..I. LLPMMSISAL TAL..F.... ..V...M... ...V..L...
 7    ...L...... IQ..AT.AFV LLS.MTVAIL TSIV.F...L ......M... ...V..L...
 8    ...L...... IQ..AT.VFV LLPMMTVAIL TAAV.F...L ..V...M... ...V..L...
 9    ...L...... IQ..AT.VFV LLPMMTVAIL TA.V.F...L ..V...M... ...V..L...
10    ...L...... .Q.VSMT.FV .IPVISISI. TSLL.L.... ..L...V... ...I..LT..
11    A..L...... IF.LST.IWL L.SSLMISVP .LI.F...FV ...G..C... ....A.IHF.
```

(Figure 5-1 continued)

```
                ← ATP6 || COIII →
  1    LHDNTHQSHA YHMVKPSPWP LTGALSALLM TSGLAMWFHF HSMTLLMLGL LTNTLTMYQW
  2    .......T.. ....N..... .......... ....T..... N......I.. T..M......
  3    .......T.. ....N..... ....F....L ....V....Y N.I...T... ...I......
  4    .......T.. ....N..... .........L ....V....Y N.TI..S... ...I......
  5    .......T.. ....N..... .........L ....I....Y N.S..MFM.. T.ML......
  6    .QE.A..A.S ....D..... IF..AA...T ....I....Y S.T...TM.. .SML.V.L..
  7    .QE.A..A.. ....D..... ....VA...L .......... G..I..T... I.MV...I..
  8    .QE.A..A.. ....D..... ....IA.... .....I.... ..T..MT... ILLL......
  9    .QE.A..A.. ....D..... ....IR..FL .....I.... Q.V...T... ILLL......
 10    .QE.S..A.. ....D..... ....GA.... .........K N.CI.MT... ILML......
 11    .QQ.AI.-.P ..L.DQ.... .D..F.G.M. ...NVL...T QKTN.TLV.F .LLITN.VN.

  1    WRDVTRESTY QGHHTPPVQK GLRYGMILFI TSEVFFFAGF FWAFYHSSLA PTPQLGGHWP
  2    ....I....F ......A... .......... I...L..T.. .......... ...E...C..
  3    ....I..G.. ......I... .......... V.......... .........V ..HD...C..
  4    ...II..G.. ......I... .......... V.......... .........V ..HD...C..
  5    ...II..G.F ......V... .......... L......I.. .......... ...E...C..
  6    ....V....F ......T... .......... ...A...L.. ....F..... ...E...Q..
  7    ....I..G.F .......... .......... .......I.. .....N.... ..YE..EC..
  8    ...II..G.F .......... .......... .......L.. .......... ...E...C..
  9    ...II..G.F .......... .......... .......L.. .......... ...E...C..
 10    ...IV..G.F L....S...Q .......... I...C..... ......A... ...E..LT..
 11    ...II.KANF ..S..AI.N. .M........ ....C..FA. ....F..... .SVEI.VA..

  1    RTGITPLNPL EVPLLNTSVL LASGVSITWA HHSLMENNRN QMIQALLITI LLGLYFTLLQ
  2    P...H..... .......... .......... ......GD.K H.L...F... T..V......
  3    P...S..... .......... .......... ......GK.. H.N....... M......I..
  4    P.......... .......... .......... ......G... H.N....... .......I..
  5    P...H..... ........I. .......... ......G..K .......... S......I..
  6    P..VK..... .......AI. .....TV... ...IT.G..K .A.H..TL.. ...F...A..
  7    P........F .......A.. .....TV... ...I.HGD.K EA..S.TL.. .......A..
  8    P..M...D.F .......A.. .....TV... ...I..GE.K .A..S.AL.. .......A..
  9    P......D.F .......A.. .....TV... ......GA.K .A....AL.. I..V...A..
 10    P...N....F .......A.. ......V... ...IT.K..T ETT...TL.V .......A..
 11    PS.......F L......G.. .S...TLS.S ...ILAG..T ES....FL.V A..S...A..

  1    ASEYFESPFT ISDGIYGSTF FVATGFHGLH VIIGSTFLTI CFIRQLMFHF TSKHHFGFEA
  2    ....Y.A... ....V..... .......... ........IV ..F...K... ..N.......
  3    ......TS.S .......... .M........ ........IV .LL...K... ..........
  4    ......TS.S .......... .M........ ........IV .LL...K... ..........
  5    .M..Y.AS.. ....V..... .......... ........IV .LL...FY.. ..T.......
  6    .M..H.AS.S .A.SV..... .......... .....S...V .LL.LIK... .PN.......
  7    .M..Y.A... .A..V..... .......... .....L..SV .LL..IQY.. ..........
  8    .M..Y.A... .A..V..... .......... ........AV .LL..IQY.. ..E.......
  9    .M..Y.A... .A..V..... .......... .....S..AV .LL..IQY.. ..E.......
 10    IM..Y.T... MA..V..... .......... .....L..LT .LL.H.QY.. ..........
 11    .W..IDA... .A.SV..... .........Q ....T...MV .LF.TAGR.. STH.......
                        ← COIII || ND3 →
  1    AWYWHFVDVV WLFLYVSIYW WGSFWLPQLN GYMESTPYEC GFDPMSPARV PFSMKFFLVA
  2    .......... .......... .......... V.S.TS.... .....GS..L ..........
  3    .........I .......... ........M. L.S.AN.... ....T.S..L ..........
  4    .AWY...... .......... ........M. L.S.AN.... ....T.S..L ..........
  5    .......... .......... .........Y L.L..S.... ....LGS..L ..........
  6    ........II .....M.M.. ......A.MA PDT.LS.... ....LGS..L ...IR.....
  7    .......... .......... ........MT PD..LS.... ....LGSM.L ....R...I.
  8    .......... .......... ........M. PDA.LS.... ....LGS..L ....LR.....
  9    .......... .......... ........M. PDA.LS.... ....LGS..L ...IR.....
 10    .......... .....I.... .......IMK PDS.LS.... ....QGS..L ...LR.....
 11    .......... .FV..WL... ..AALPNRTS DSEK.S.... ....LNS..L ...FR.....
```

(Figure 5-1 continued)

```
                                                                  ←ND3 | |ND4L→
  1  ITFLLFDLEI ALLLPLPWAL QTTNLPLMVM SSLLLIIILA LSLAYEWLQK GLDWTEGMLV
  2  .......... .........S ..A..NT.LT MA.F...L.. V......T.. ..E....L.M
  3  .......... .........I ..IKTST.MI MAFI.VT..S .G.....T.. ..E....T.M
  4  .......... .........I ....TTT.MA TAFI.VT..S .G.....T.. ..E....TFM
  5  .......... .........I .LPSPFTTLI L.YC..ML.T VG.....I.. ..E....V.I
  6  .L........ .........I .LAHPMMTLT WATTI.AL.T FG.I...T.G ..E.A..LAF
  7  .L........ .....F...A .LNTPSIVIL WAA.ILTL.T .G.I.....G ..E.A..LAL
  8  .L........ ........GD .LH.PTGTFF WATTVL.L.T .G.I...T.G ..E.A..LAF
  9  .L........ ....A...GD .LYSATGTFF WATAVL.L.T .G.I...T.G ..E.A..LAF
 10  .L........ .....S...T NIS.PEFTLL WAS.FVLL.T .G.I.....G ....A..LSL
 11  .L........ ...F...A.S LI.PPSTLIP I.MVFMV..T .G.VF..ING ..E.A..I.L

  1  YRSHLMSSLL CLEGMMLSLF IMATLMTLNT HLLANIVPIA MLVFAACEAA VGLALLVSIS
  2  .......... .......... V..A.TI..S .T..SMM..I L......... L..S...MV.
  3  F......T.. .....V.... ..TSVTS..S NSMSSMPIPI T......... .......KV.
  4  F......T.. .......... V.TSTS...S NSMISMTIPI T......... .......K..
  5  .......T.. .......... .FMAA.ITHF .FSISMM.LI L...S....G ..........
  6  H.T..I.A.. ...S....M. .PLSIWPVEN QPSFAL...L ..A.S....G T...M..ASA
  7  N..PIL.I.. .....L.MSM DGIV.TP.HL TY.SSMMLYI ..P...P... T..S.NSDHY
  8  H.T..L.A.. .......... .ALA.WA.QF ETGFSTA.ML L.A.S....S T......ATA
  9  ..T..L.A.. .......... .ALA.WA.QF ETGFSTA.ML L.A.S....S A.PG...ATA
 10  Q.K..L.L.. T..S.A.A.Y VSTA.WA..N T.PIMAA.LI I.T.S....G M..S.MIATA
 11  N.L.FL.I.. ...LLLI... .GIAIWNN.. GVPQ.TTFNL F.TLV....S I..S.M.GL.
            ← ND4L || ND4 →
  1  NTYGLDYVHN LNLLLKLIVP TIMLLPLTWL SKKHMIWINT TTHSLIISII PLLFFNINNN
  2  ....T...Q. ......Y.I. ....M..... ..NN...V.S .A...L..FT S..LM.FGD.
  3  ....T...Q. ......I.L. SL........ .SPKKT.T.V .SY.FL..LT S.TLLWTDE.
  4  ....T...Q. ......I.F. S......... .ANKK..T.V .SY.FLV.LL S.SLLWQDE.
  5  ....N.Q.Q. ......ILL. .L..I..... ..NKWL.... ..Y..L...T S.PMLYHPMD
  6  R.H.S.HL.. ......I.L. ......TAL. .PAKSM.T.. .MY..L.AS. S.HWLTPSYY
  7  T.H.T.KLFS ......ILL. .L..I.S... TN.KWL.PSL .SQ.....LL S.MW.FQSET
  8  R.H.T.RLQ. ......VLI. ....F.TI.. TSPKWL.TT. .A...L.AS. S.TWLKTSET
  9  R.H.T.RLQ. ......VLI. ....F.TI.. TPPKWL.ST. .A.G.L.AL. S.TWLKSSEV
 10  R.HNT.QLKA ........I. S...I.M.F. IN.KSLLWTA ..FFSFLIAA LSTLTLMDVA
 11  R.HSSNL.GS .S..ILFT.G MATTTL.IPS N.LWAGA.FQ SALLSLL.L. V.NNHWTASW

  1  LSCSPTFSSD PLTTPLLMLT TWLLPLTIMA SQRHLSSEPL SRKKLYLSML ISLQISLIMT
  2  SNF.LL.F.. S.S....I.. M.....ML.. ..H...K.N. T....FIT.. ....LF....
  3  YNF.NM.... ..S...II.. A.....ML.. ..N..KKDNN VLQ...I... .....L....
  4  YNF.VM.... ..S...II.. ......MML. ..N.MKK.NM MHQ...I... .....L....
  5  .GNFNNSF.. S.SS...V.S C.....M... ..N..NK.S. M......T.M VI..S...IA
  6  PKTLTLWTG. QIS....V.S C.F...M... ..G..QH..H K..RMFI.T. .II.PFI.LA
  7  THF.NYLMT. QIS....I.. C.....MLI. ..N...N..I ..QRTFIT.. VF..L....A
  8  GTS.NMYLAG ..L....V.. C.....M.L. ..N.INP..I ..ER..ITL. AL..TF...A
  9  GAT.SLYLA. ..S....G.. C.....MVL. ..N.ITP..I I.QR..ITL. A...TF...A
 10  ENSTNPLL.. QFSC..I..S C........G ..A.MKT..I T.Q.TMI.L. .L..VL.CI.
 11  HNL.SILA.. TISA..II.S C..A.IALI. .KGQ.NNSSD LGSRVFII.I .VITGA..I.

  1  FTATELIMFY IFFETTLIPT LAIITRWGNQ PERLNAGTYF LFYTLVGSLP LLIALIYTHN
  2  ...M...L.. .L..A..V.. .I........ T......L.. .....A.... ..V....IQ.
  3  .S........ .L..A..... .I........ T......I.. .....I..I. ......LIQ.
  4  .S.....L.. .L..A..... .I........ T......I.. .....I..I. ......SIQ.
  5  ..SS..M... .L........ .I........ N......I.. .......... ..V..LTMNK
  6  .S....ML.. .S..A..... .IL....... ....S..I.L .....IS... ..VSIL.L.T
  7  .S.....L.. .M..I..... .I........ A......... .....A.... ..V..LSLYS
  8  .G...I.... .M..A..... .I........ T......... .....A.... ..V..LLLQQ
  9  .G...I.... .M..A..... .I........ T......... .....A.... ..V..LLLQQ
 10  .G.SN.L... .A........ .L........ K...T..L.. .....SA... ..L...MIQT
 11  .SSL...L.. .VV....... .IL.....A. M..CQ..L.. M....F.... ....I.AIYI
```

(Figure 5-1 continued)

```
 1    TLGSLNILLL TLTAQELSNS WANNLMWLAY TMAFMVKMPL YGLHLWLPKA HVEAPIAGSM
 2    .V....F.M. QYWV.PVH.. .S.VF....C M......... .......... ..........
 3    HV.T..LMI. SF.THT.DA. .S...L...C M...LI.... ..V....... ..........
 4    SM.T..F.I. S..THP.PST .S.TIL...C M....I.... ..V....... ..........
 5    N..T.H..MN SILINQ.NYT LS.STL.Y.C MT...I.... .......... ..........
 6    NT.T.HLPII K..HPN.PA. .TSL.SS..L L......A.. .......... ..........
 7    ST.T.SLN.. Q.LPNHIPMT ...YSW...C LL......... ..T....... ..........
 8    ST.T.SM.V. QYSQPLQL.. .GHMFW.AGC LI..L..... ..V....... .....V....
 9    .N.T.SL.I. QHSQPLALT. .GHKIW.AGC LI..L..... ..V....... .....V....
10    H.N..S.YII P.SNLT.LTP .SET.W.I.C FL..LI.... .IF....... ..........
11    SSS..S.PNV N.LWANDGSI ESLTMW.ALS INC.FNNL.V ..F....... .....V....

 1    VLAAVLLKLG GYGMMRLTLI LNPLTKHMAY PFLVLSLWGM IMTSSICLRQ TDLKSLIAYS
 2    .......... ....L.I... ...M.DF... ..IM...... .......... ..........
 3    I...I..... S...I.ISI. .D....Y... ..IL...... .......... ..........
 4    I...I..... ......VSI. .D....SL.. ..II...... .......... ..........
 5    ....I..... ...I..IS.F TE.M.M.LL. ..II..M... .......M.. ..........
 6    L...L..... ...I..V..L ME.VSNFLH. ...T.A...A L......... ..........
 7    ....I..... ...II.ISIT .S.SM.EL.. ...I.....I .......... .....M....
 8    .......... ......MMVM .D..S.EL.. ..II.A...I ...G...... ..........
 9    .......... ......MMVV .D..S.EL.. ..II.A...I ...G...... ..........
10    I...I..... ....I.MSSL FI....DL.V ..MIIAM... .V....... .....M....
11    I...I...I. ...L...IAL FSTISMNLSL ALI.FCT..A LI..V..V.. ....A.....

 1    SISHMALVVT AILIQTPWSF TGAVILMIAH GLTSSLLFCL ANSNYERTHS RIMILSQGLQ
 2    .V......IV .........Y M..TA..... .....M.... .......I.. .T...AR...
 3    .V......IA S.M....... M..TM..... .......... .......I.. .T..MAR...
 4    .V......I. ..M....... M..TM..... .......... ..T....I.. .T..MAR...
 5    .V......II .A...STT.. M..T...V.. .....M.... ..T....I.. .T...AR...
 6    .V...G..IA .SM...Q... S..M....S. .......... ..T....... ..L..TR...
 7    .V...G..IS .GNN...KAL ...M..NTSD ...H.A.C.. .KQS...... .ALL..R..E
 8    .V...G..AG G.......G. S..I...... ..V..A.... ..TA...... .T...AR...
 9    .V...G..AG G.......G. ...I...... ..V..A.... ..TA...... .T...AR...
10    .V...G...A G.FTM...AW S..LAM.... ..V..G.L.. ..IT.....T .SIFMNR..K
11    .VG..SI.AA ..FSE.S.GM N..LM..V.. ..V..A..S. ..TV...SGT .TLAITR..K

 1    TLLPLMAFWW LLASLANLAL PPTINLLGEL SVLVTTFSWS NITLLLTGLN MLVTALYSLY
 2    .......T.. .....T.... ......I... F.VMS..... ...II.M.V. .VI.......
 3    MVF....T.. .M........ ..S...M... FITMSL.... .F.II.M.I. III.GM..M.
 4    MIF....T.. .......... ..L...M... FIVMA..... .PSII.MAT. IVI.GM..M.
 5    LI....TT.. .T........ .......... MIITAS.... .FSI..L... TVI.......
 6    P.....SV.. ...N.T.M.. ...T..MA.. TIM.AL.N.. SP.II...TA T.L..S.T..
 7    .I....GT.. .ISN...M.. ..SP.WM..I TIMTAL.N.. SW.II..D.G T.L..S....
 8    VIF..T.V.. FI.N...... ..LP..M... MIIT.L.N.. PW.I.....G T.I..G....
 9    MIF..T.V.. FI.N...... ..LP..M... MIIT.L.N.. PL.II...TG T.I..G....
10    ..F...S... .MMTF..M.. ..FP.FMA.I LIITSL.N.. .W.I..L..S .TL...F..N
11    L....STL.. ..MCA...G. ..SP..I..I LI.SSLI... VWLFPIV.FA QVFG.I...M

                                       ← ND4 || ND5 →
 1    MFTTTQWGSL THHNNMKPSF TRENTLMFMH LSPILLLFII SLFPTTMFMC LDQEVIISNW
 2    .LIM..R.KY .Y...IS... ....A..SL. IL.L.....T .MI..M..IH SG..L.....
 3    .II...R.K. .N.I.LQ..H ...L...AL. MI.LI..... ..L.LL..FH NNM.YM.TT.
 4    .II...R.K. .S...LQ..H ...L...AL. II.LM..T.N ..L.LLL.FH HNT.YM.T..
 5    .L..S.R.KF ...TSLY... ...HM..TL. IM.LI...MM ..PSLLL..Y KG..S..T..
 6    .LLS..R.T. PS.TTTPN.N ...HL..TL. II.M.T..L. ..I...I.IH SGA.S.ATH.
 7    ..LM..R.MT PE..AIN.TH ...H...T.. .I..IP..L. .II.LII.LD QGL.S.TT.F
 8    ..LMS.R.PT PN.MGLQ.FH ...HL..TL. .I.VI...FV ..L.LMI.LN .KT.G..T..
 9    L.LMS.R.PT PK.VGLP.FH ...HL..AL. .I.V....L. ..L.LAV.LD QGT.T.VT..
10    .LIM..HEHP NK.APVN..T ...HL..L.. MA..I..MF. ..I.L.IYLN ENM.TTLTMK
11    I.QLS.Q.TP FTSIINVF.. S..HLFAAL. IL.LI.IAFL .VLSLLVTCN NSIQS..TLS
```

(Figure 5-1 continued)

```
 1    HWATTQTTQL SLSFKLDYFS MMFIPVALFV TWSIMEFSLW YMNSDPNINQ FFKYLLIFLI
 2    ..L.I..LK. .....M.... .......... ........M. ..Y......K ......L...
 3    ..V.MNSME. KM...T.F.. IL.TS..... .....QL.S. ..H......R .I...TL...
 4    ..L.INSIK. TM...I.... IL.LS.S... .....Q..S. ..H...H..R .I...MM..N
 5    ..FSIHSFNI .M...M.F.. II...I.... ..A.L..... ..H.....S. .....I...L
 6    E.QFIPNFKI PI.L.M.MY. ...F.I.... ....L..AT. ..A.E.F.TK ..T...T...
 7    ..MNIN.FDI NM...F.IY. SI...I.... ....L..AT. ..A...M.SR ......T..V
 8    Q.MN..AFDV NI...F.HY. LI.V.I..Y. ....L..A.. ..H.....DR ......T..V
 9    ..MN.TMFDI NI...F..Y. LV.T.I..Y. ....L..AS. ..HA..YV.R ......T...
10    P.MDWALFNI A....I.KYT VI.T.I..MI ........Q. ..AKERHMDK ......L...
11    L.LSNTPLNI ..N.IY.QYF LV.LS...I. ........FY ..TE...SSA ..RL.T...L

 1    TMLILVTANN LFQLFIGWEG VGIMSFLLIS WWYARADANT AAIQAILYNR IGDIGFILAL
 2    .......... .......... .........G ...G...... ..L....... .........M
 3    .....TS... M......... .........G ...G.T.... ..L....... .........M
 4    N....TS... .......... .........G ...GL..... ..L....... V........M
 5    ..I...S... .......... .........G ...G.S.... ..L....... ......M.TM
 6    A..T.TI... M.L..V.... .........G ..QG..E... ..L..MI... .....L..SM
 7    A.V....... F..F...... .........G ......EP.. ..L..VI... V....L..SM
 8    A.I....... M......... .........G ..HG...... ..L..VI... V....L.MTM
 9    A.IT...... M......... .........G ...G...... ..L..WI... V....L.MSM
10    ..ITFIS... .L........ .......... ..SG.TK..I S.L..VA... .....LMMSM
11    N....TCS.S ..LI.L.... G.FL...... ..TT.N..SS S.LE.VIT.. ..N..L.TFM

 1    AWFILHSNSW DPQQMLNANS LTPLLGLLLA AAGKSAQLGL HPWLPSAMEG PTPVSALLHS
 2    ...LTNL.T. .L..I..PS. NM..I..A.. .T.....F.. .......... ..........
 3    V..S.NM... EL..IS.N.N .I..M...I. .T.....F.. .......... ..........
 4    T..C.NM... EL..I..N.N .V..T...I. .T.....F.. .......... ..........
 5    ..LM.NC... .L.HISMNMH PIA....II. .T.....FS. .......... ..........
 6    ..LASSL.T. EI..ITHP.T PL.....I.. .T.....F.. .....A.... ..........
 7    ..VAMNL... EM..V..SDN .L.....I.. .T.....F.. .....A.... ..........
 8    ..LAMNL... EI..I.SK.D MI..M..A.. .T.....F.. .....V.... ..........
 9    ..LAMNL... EI..I.SKDD ML..M..I.. .T.....F.. .....A.... ..........
10    V.MCSNT... .L..I.LSDQ YI.T..F.I. .T.....F.. .....A.... ..........
11    .LSA.NF..S NLTNI.SS.E N....PF... .......F.. .....ALL.. ..........

 1    STMVVAGIFL LIRFHPLAEN SPLIQTLTLC LGAITTLFAA VCALTQNDIK KIVAFSTSSQ
 2    .......... ....Y..T.. NKY..SI... ........T. M......... ..I.......
 3    .......... .V.....TT. NNF.L.TM.. ...L....T. I......... ..I.......
 4    .......... M......TS. NST.M.AM.. ........T. I......... ..........
 5    .......... ......ML.. NKTML.I... ...L....T. M..IM..... ..........
 6    .......... ...T..FLSS NKTAL.TC.. ...LS..... T......... ..I.......
 7    .......... ...IS.MMN. NQTAL.IC.. ...M....T. A......... ..........
 8    .......... ...L..RM.. NQ.AL.TC.. ...L.S..T. T......... ..........
 9    .......... ...L.AIM.. NQ.AL.TC.. ...L.S..T. A......... ..........
10    .......V.. ...L...FQ. Y..MLEM... ...M..IC.. L..T...... ..I.......
11    .......V.. .V.TSE.FSS PLITHS.V.I ..GT.A.... ST.IA.H... ..I.Y..T..

 1    LGLMMVTIGI QPHLAFLHIC THAFFKAMLF MCSGSIIHNL NNEQDIRKMG GLLKTMPLTS
 2    .......... ..Y....... .......... ........S. .D........ ..F.A..F.T
 3    .......L.M .......... .......... ........S. AD........ NIT.I..F..
 4    .......L.. ..Y....... .......... ........I. .D........ NMM.A..F..
 5    .......V.L .......... .......... L......... .D........ ..FY.L.I..
 6    .........L L.Q......S .......... L...L...S. .G........ C.Q..L.M.T
 7    .........L IFQ...F... NN....VYY. F...QYSSC. .D........ ..QNSL.I.T
 8    .........L ..Q....... .......... L.......S. .D........ ..FNI..A..
 9    .........L ..Q....... .......... L.......S. .D........ ..QNLL.F..
10    ......AV.L H..I....M. .......... L.........M ........FS C.NNNL...T
11    ....VTA... ..A...F... .......... L....V..S. SD...L.... ..S.LL.V..
```

(Figure 5-1 continued)

```
 1    TSLTIGSLAL AGMPFLTGFY SKDHIIETAN MSYTNAWALS ITLIATSLTS AYSTRMILLT
 2    .A.IV..... T......... ...L...A.. K........L M......F.A I....I.FFA
 3    SC.V...... T......... ...L...AI. TCN......L .......M.A M..M.I.YFV
 4    SC.I...... T......... ...L...AI. TCN......M .......M.A V..M.L.YFV
 5    SA.MT..... M.T...A... ...S...AM. T....S...T .........A I..L.I.YY.
 6    SC....N... M.T...A... ...L...NL. T..I.T.... L..L...F.A T..L..T..V
 7    SC........ T.T...A..F ...A...AL. T.Q..T...T L......F.A I..F.V.FFA
 8    .YF....... T.T...A..F ...A...AL. T..L.....T L......F.A V..F.LVYFV
 9    .C........ T.T...A..F ...A...AM. T..L.....T L......F.A V..F.VVFFV
10    .CM....A.. M.L...A..F T..L.L.AL. T........M V..M.VT..T ...S.L.IMS
11    SC.IL..... MA-.L.A... ...L.L.ATS A.VL.LLGIV LSIV..M..A V..F.I.FFC

 1    LTGQPRFPTL TNINENNPTL LNPIKGLAAG SLFAGFLITN NISPASPFQT TIPLYLKLTA
 2    .L........ V.......L. I.S..R.LI. .....YI.S. ..P.TTIP.M .M.Y...T..
 3    TMTK....P. IS....D.D. M....R..F. .I....V.SY ..P.T.IPVL .M.WF..T..
 4    TMTK..YSP. IT......N. I....R..L. .IL.....SL ..P.TNIQFL .M.DN..M.S
 5    .L.H...M.M SPL.....N. I...IR..L. TI....ML.T .MP.SYSITM .M.MFI.QM.
 6    Q..HT.T.SN HP....T.PA IL..MR..L. .IM..L..SS L.L.PKTPPM .M.TIT.TA.
 7    SM.H..SNP. SP.....K.V I....R..W. .IV..L..AS .ML.INSPIM .M.TLA.QA.
 8    IM.T...LP. SP......LV I.T..R..W. .II..LI..Q .FP.MKTPIM .MSIT..MA.
 9    SM.T...LP. SP......AV I....R..W. .II..LI..S .LL.SKSPVM .M.PT..MA.
10    AS.T..YLP. .PTH...FI- K..L.R..W. ..IS.LIL.S TLP.MK.QIF .M.T.I.TI.
11    FSLS.SCSSP FSHS.E.FN. N.ALLR..T. TIAS.WFFS. LLFAPPS.NV .SLAKGTPLI

                                                  ← ND5 || Cyt b →
 1    LAVTFLGLLT ALDLNYLNCT FYFSNMLGFY PSITHRTIPM RKINPLMKLI NHSFIDLPTP
 2    .I..I..FIL ..EISNMKNA .K..TL..YF .T.M..LA.I ..SH....IV .NA.....A.
 3    .IISV..F.I ..E..N.MPY SS..TL...F ...I..IT.. ..TH..F.I. ........A.
 4    .INYN..FAI ..E..N.TKQ SS..TP..Y. .P.M..I... ..SH..F.I. ........A.
 5    .M..TT..MM GME..S..H. NN.LT..... TQ.M..MQ.I ..TH....I. .D........
 6    II..T..IIL ..E.SS.SPL MN..SS..YF NPL...IS.I ..SH..L.M. .N.L....A.
 7    II.SVT..II .M..SK.TNI HS...L...F .T.I..MM.I ..SH..I.I. .N........
 8    .M..IA...V .ME.AN.SS. HH........ .M.I..L..L ..TH..I.IA .DALV.....
 9    .M..AI..F. .ME.AT.SQ. HH.......F ..VV..LMKL ..TH..I.IA .DALV...A.
10    .MMFIIS.II SME.TNKIT. .S.FTQ.A.. .H.I..LTSI ..TH..LS.G .SILV...S.
11    VPIIGVAA.F MSLISSTSNS IGSNAHSATT SQWFFVDAVL ..EH.IFRIL .ST.V...L.

 1    SNISAWWNFG SLLGACLILQ ITTGLFLAMH YSPDASTAFS SIAHITRDVN YGWIIRYLHA
 2    ....S..... ....I..... .L........ .TS.TT.... .VT..C.... .......M..
 3    ....S..... ....V..MV. .I........ .TS.TM.... .VT..C.... ...L...M..
 4    ....S..... ....V..MV. .L........ .TS.TM.... .VT..C.... ...L....Q.
 5    .......... ....V...I. .L........ .TS.TL.... .V...C.... ...L..NI..
 6    .......... ...AV..MT. .L...L.... .TA.T.L... .V..TC.N.Q ...L..N...
 7    ....SL.... ....V...A. .I........ .TA.T.M... .V...C.... ...L..N...
 8    .......... ....L...T. .L........ .TS.I..... .VT..C.... ...L..NV..
 9    ....V..... ....L...T. .L........ .TS.I..... .V...C.... ...L..NI..
10    A......... ...SL..... .I...I.... .TANTEL... .VM..C.... N..LM.N...
11    ..L.I...S. ....L..VV. .L..I..... .TA.ITL... .VM..L.... ...FL..V..

 1    NGASMFFICL FLHIGRGLYY GSFLYSETWN IGIILLLATM ATAFMGYVLP WGQMSFWGAT
 2    .......... YM.V...... ..YTFL.... ..V....TV. .......... ..........
 3    .......... ...V...... ..YTFM.... ..VL..F.V. .......... ..........
 4    .......... ...V...... ..YTFL.... ......F.V. .......... ..........
 5    .......M.. ...V...I.. ..Y..K.... ..V....TV. ....V..... ..........
 6    ....F....I .......... ..Y..K.... T.V....TL. ....V..... ..........
 7    ....F....I Y......... .....K.... ..V...FLV. ....V..... ..........
 8    ....F....I YM..A..... ..Y..K.... ..VV...LV. M...V..... ..........
 9    ....F..... Y...A..... ..Y..K.... ..VV.F.LV. M...V..... ..........
10    .........I YA.....I.. ..Y..K.... V.V..FAL.A ....V..... ..........
11    ..V.L....M YC........ ..YNKI.... V.V..F.V.I L........V .......A..
```

(Figure 5-1 continued)

```
 1    VITNLLSAIP YIGTDLVQWI WGGYSVDSPT LTRFFTFHFI LPFIIAALAT LHLLFLHETG
 2    .......... ....N..E.. ...F...KA. .....A.... .....M.I.M V.........
 3    .......... ....T..E.. ...F...KA. .....A.... .........I V.........
 4    .......... ....T..E.. ...F...KA. .....A.... .........I V.........
 5    .......... ...ST..E.. ...F...KA. .....A.... .....L.MVV V.........
 6    .....F.... ...HT..E.A ...F...N.. .....AL..L ...A..GITI I..T....S.
 7    .......... ...NV..... ...F...NA. .....A...L ......GASI ..........
 8    ........V. .M.DM..... ...F...NA. .....A...L ...V...ATI I.........
 9    ........V. .V.DM..... ...F...NA. .....A...L F...V..VTI ..........
10    .....I..M. .V.N.I.V.L ...F..SNA. .......... ....L..MTM I.IM...Q..
11    .....V.... ....II...L ...F...NA. .....P...L F........V I..V...NS.

 1    SNNPLGITSH SDKITFHPYY TIKDALGLLL FLLSLMTLTL FSPDLLGDPD NYTLANPLNT
 2    ....T..S.D V...P..... ....I..A.. LI.A..L.V. .A........ ...P......
 3    ....T.LN.D A...P..... ....I..I.I MF.I....V. .F..M..... ..MP......
 4    ....T.LN.D A...P..... ....L..VFM L..F....V. .F........ ...P......
 5    .S..T.LDPN ....P..... .M..I...F. MIII.LS.AM .......... .F.P......
 6    .......S.D ....P..... SF..I...T. M.TPFL..A. ...N.....E .F.P....V.
 7    .T..T.LN.D P..VP....F SY..L..F.I M.TA.TL.AM ...N...... .F.P....I.
 8    ....I.LN.D A..VS....F SY..L..FVI M..A.TL.A. ...N.....E .F.P....V.
 9    ....A.LN.D A...S....F SY..L..FVV M..G.T..A. ...N.....E .F.P....V.
10    .S..M..N.N L...Q....F SF..I..FVI L.GI.FMIS. LA.NA..E.. .FIY....S.
11    A...FAFN.N Y..AP..I.F .T..TV.FI. LVAA.FS.A. LF.GA.N..E .FIP....V.

 1    PPHIKPEWYF LFAYTILRSV PNKLGGVLAL LLSILILAMI PILHMSKQQS MMFRPLSQSL
 2    .......... ....A....I .......... AF......L. .L..T...R. ........C.
 3    .......... ....A....I .......... I.......LM .F..T...R. L....IT.I.
 4    .......... ....A....I .......V.. I.......FL .F..T...R. LT...IT.I.
 5    .......... ....A....I .......... .A...V.LI. .M..T.T.R. .A...I..T.
 6    .......... ....A....I .......... AA.V...FL. .F..K...RT .T......T.
 7    .......... ....A....I .......... V.......LM .L..T...R. L....FT.IM
 8    .......... ....A....I .......... .F...V.MVV .L..T...RG LT...IT.F.
 9    .......... ....A....I .......... .F...V.MVV .V..T...RG LT...AT.F.
10    .......... ....A..... .......V.. AAA.M..LI. .FT.T...RG .Q....A.IT
11    ....Q..... ....A....I .......I.. VAA..V.FLM .L.NT..NE. NS......AA

                                                    ← Cyt b |
 1    YWLLAADLLI LTWIGGQPVS YPFTIIGQVA SVLYFTTILI LMPTISL
 2    F.A.V....T .........E H.YIT...L. .....LL..V ....AGT
 3    ..I.V.N... .........E H..I....L. .IS..SI... ...ISGI
 4    ..I.V.N..V .........E H..I....L. .IS..SI... ...ISGI
 5    F.M.T.N.I. .........E Q.YIT...W. .IS...I.I. ...LAGM
 6    F...V.N... .....S...E H..I....M. .LS...IL.. .F...GT
 7    F.A.V..T.. .........E D.Y.M...L. ..I..SIFI. MF.LMGW
 8    F.T.V..MI. ......M..E H..I....I. .....ALF.. F..LAGW
 9    F.T.V..MI. ......M..E H.YI....I. .I...ALF.. .I.LAGW
10    F.I.I...AL ...L..E.AE ...ILMT.I. .TV..MIFIL VF.ILGY
11    F...V.H.F. .....S...E ..YVLL.... .....SLFIF GF.IV.S
```

RESULTS

Protein Phylogeny of Vertebrates

Three possible trees were obtained by PROTML program under the constraint that the well-established vertebrate relation of (((((((human, cow, (mouse, rat)), marsupial), chicken), frog), carp, loach), lamprey), urchin) is given. The three tentative trees were tested by three different models, JTT-F, Dayhoff-F, and Poisson-F (Table 5-1). All models provided the same phylogenetic trees, but JTT-F model has the lowest AIC value, indicating it is the most appropriate model for amino acid evolution. Thus the other two models were excluded in further analysis. By comparing the log-likelihood of each tree topology, the best tree topology was ((((((human, cow), (mouse, rat)), marsupial), chicken), frog), carp, loach), lamprey), urchin) (see Table 5-1 and Fig. 5-2). As seen in Table 5-1, the bootstrap value (94.7%) also strongly support the tree among the four candidates. The tree topology of mammals is identical to that of Cao et al. (1993)'s study. Using the best tree, I estimated the divergence rate between cold- and warm-blooded vertebrates in two ways.

Rate of Evolution among Vertebrates

The branch lengths of the two major lineages, cold- and warm-blooded

Table 5-1. Phylogenetic relationships among ten vertebrates. JTT-F is superior to other two models.

Tree topology	Model	In L	AIC	Boot P.
((((((Human,Cow),(Mouse,Rat)), Marsupial),Chicken),Frog), (Carp,Loach)),Lamprey,Urchin);	JTT-F	(-38974.4)	(78024.8)	0.947
	Dayhoff-F	-315.2	+630.5	0.884
	Poisson-F	-2762.3	+5524.6	0.703
((((((Human,(Mouse,Rat)), Cow),Marsupial),Chicken),Frog), (Carp,Loach)),Lamprey,Urchin);	JTT-F	(-39006.4)	(78088.8)	0.048
	Dayhoff-F	-307.9	+615.8	0.109
	Poisson-F	-2742.5	+5484.9	0.293
((((((Cow,(Mouse,Rat)),Human), Marsupial),Chicken),Frog), (Carp,Loach)),Lamprey,Urchin);	JTT-F	(-39017.0)	(78110.0)	0.005
	Dayhoff-F	-310.5	+620.9	0.007
	Poisson-F	-2761.9	+5523.8	0.004

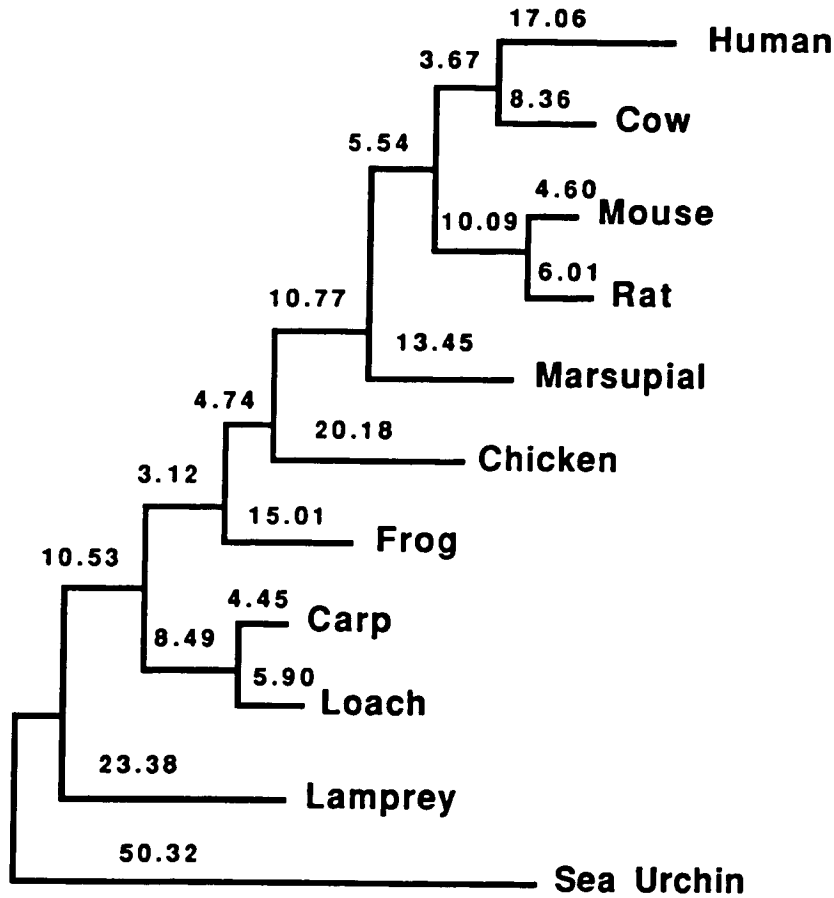


Figure 5-2. The best phylogenetic tree constructed by PROTML program. The branch length of each lineage is proportional to the number of amino acid substitutions estimated by JTT-F model. This tree is strongly supported by a bootstrap analysis (96 %).

vertebrates, were compared. The longest branch lengths of the cold- and warm-blooded vertebrates are 52.04 and 57.22 respectively (Figure 5-2). The lengths were the sums of all branch lengths consisting of the longest branches in each group. The divergence of the cold-blooded vertebrates occurred about 550 million years (myr) ago. The warm-blooded animals diverged from the cold-blooded animal (frog in this study) approximately 275 myr ago (Carroll, 1988; Colbert et al., 1991). The averaged branch lengths per million years are 0.10 in the cold-blooded vertebrates and 0.21 in the warm-blooded animals. The branch length of the warm-blooded vertebrates per million years are about twice longer than that of the cold-blooded vertebrates. It is noteworthy that the branch lengths of the two bony fishes and cow are the shortest among cold- and warm blooded vertebrates respectively, while lamprey and chicken have the longest branch lengths in the two groups. In addition, a faster rate in the rodent lineage is not evident in this study.

The number of amino acid substitutions between species was calculated in order to compare the rates of protein evolution among the major vertebrate lineages (Table 5-2). The sequence differences between species were plotted versus the divergence time in Figure 5-3. The divergence times are based on vertebrate paleontological records (Carroll, 1988; Colbert et al., 1991). According to the observed pattern of amino acid substitutions in Figure 5-3, it appears that the rate of evolution in the warm-blooded vertebrates is faster than in cold-blooded vertebrates. This supports the results from previous studies, but with a smaller difference in rate.

Table 5-2. Distance matrix based on the number of amino acids different among 11 vertebrates. All gap sites were removed from the subset data. The total number of amino acids used is 3227. The OTU labels are as follows: 1, human; 2, cow; 3, mouse; 4, rat; 5, marsupial; 6, chicken; 7, frog; 8, carp; 9, loach; 10, sea lamprey; 11, sea urchin

OTUs	1	2	3	4	5	6	7	8	9	10	11
1		671	777	797	853	1052	1043	1024	1034	1196	1471
2			640	667	715	985	924	914	921	1153	1411
3				311	778	1041	986	963	959	1141	1447
4					798	1059	1006	974	983	1163	1442
5						965	933	917	929	1155	1431
6							902	869	870	1155	1451
7								724	740	1083	1433
8									299	1021	1387
9										1036	1397
10											1454

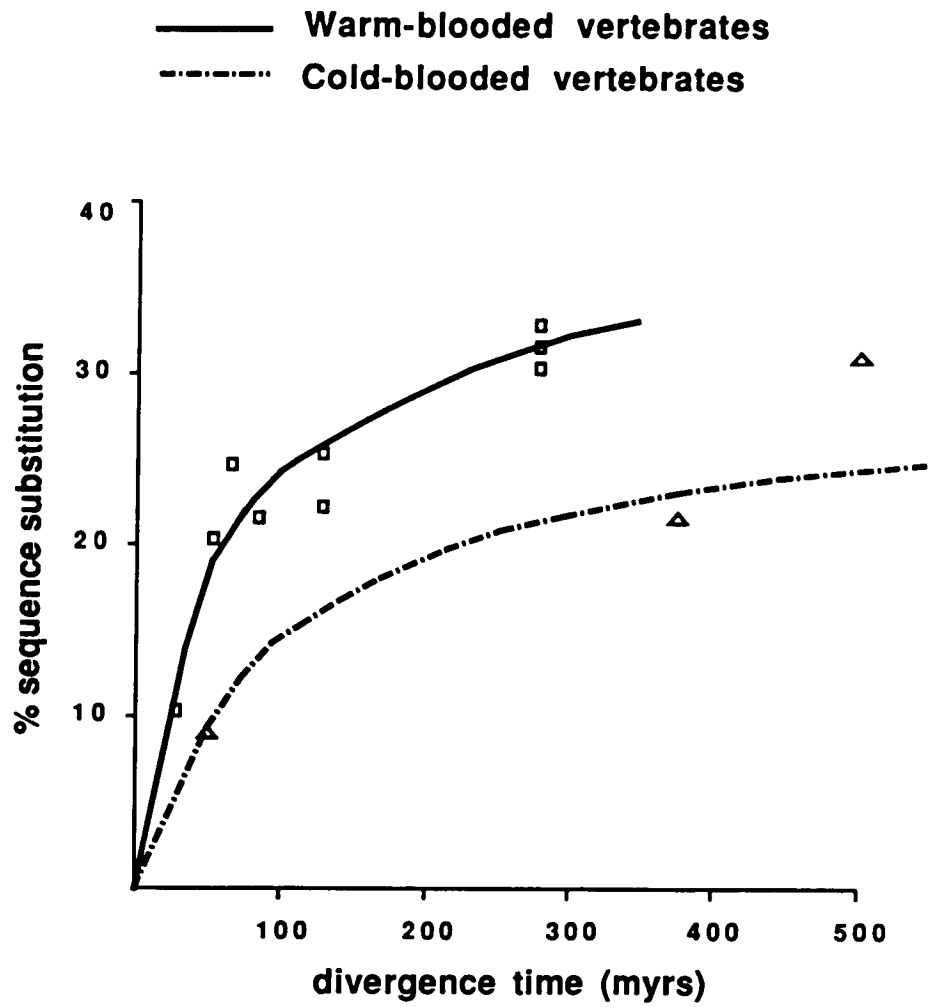


Figure 5-3. The pattern of amino acid substitutions observed in mitochondrial protein genes. The rectangles and triangles represet the warm-blooded and cold-blooded vertevrates respectively. The overall patterns are shown by two different lines.

The two recently diverged groups provide an opportunity for closer comparisons. The murids have been diverged for about 10-20 million years, and the cyprinids for about 20-50 million years. Thus it is concluded that the mammals (0.005-0.01% per myr) have accumulated sequence variation about 1.25-5 times faster than the fishes (0.002-0.004% per myr). If the averaged divergence times are considered, the former has a rate only about 3 times faster than does the latter. An important feature found from this study is the apparent saturation pattern of amino acid substitutions. In warm-blooded animals, protein sequence substitutions accumulate nearly linearly for the first 100 million years. However, the rate of observed substitutions is greatly reduced after about 100 myr. For cold-blooded vertebrates, despite insufficient data between 50 and 375 myr, it is obvious that the rate of observed variation per million years between fish and frog is much lower than between the two fishes. Thus it is evident that this reduced rate after a certain time point must be due to the multiple substitutions of amino acid sequence at the same sites.

DISCUSSION

In a previous study, Cao et al. (1993) said that the JTT-F model was the most appropriate for evolution of amino acid sequences. They used only highly conserved regions where the sequence alignments are not ambiguous. In this study, from AIC values, the JTT-F model is also superior to Dayhoff-F and Poisson-F models despite including relatively variable sequence regions, suggesting that the sequence alignment in this study is appropriate (Table 5-1, and Fig. 5-1). Only the ends of sequences that were extremely variable were excluded in this study, not the sequences from the middle of genes.

The variability of amino acids along genes is suggested to be strongly related to the secondary structure of the proteins and chemical properties of the domains (Adkins et al., 1994). Variability is also apparent from gene to gene. For instance, COI is the most conserved gene, while ATP8 has the most variable sequence, which suggests that the functional constraints vary across genes. Therefore excluding variable regions will lose potentially important information and may result in shorter branch lengths than actual divergence.

Multiple Substitutions

As seen in Figure 5-3, in the warm-blooded vertebrates, the rate of sequence

substitutions becomes saturated slightly before 100 million years. At this point, none of the models for the evolution of amino acid sequence take the saturation phenomena into account. The JTT-F model employed for the phylogenetic tree in Figure 5-1 does not consider multiple substitutions either, therefore unequivocally resulting in the shorter branch lengths of the early diverged lineages. Furthermore, direct comparisons between distantly related taxa, for instance, human and frog or chicken and fish, may not be an appropriate way to estimate evolutionary rate. Thus comparisons between taxa diverged within 100 million years may be more acceptable if the saturation is not taken into account. The time point of saturation is nearly identical to that of transversional substitutions at the third positions of four-fold degenerate codon families in the mitochondrial protein-coding genes (Irwin et al., 1991). A correction method for amino acid sequences is necessary to estimate the correct evolutionary rate as well as divergence rate between distantly related taxonomic groups.

How Much Faster Do Warm-Blooded Vertebrates Evolve?

From analysis of mitochondrially encoded proteins, it has been proposed that the evolutionary rate of cold-blooded vertebrates is by far faster (at least 6-fold) than that of warm-blooded animals perhaps because of relaxation of selective constraints (Adachi et al., 1993). A number of studies have proposed various hypotheses to explain the heterogeneity of evolutionary rate. Efficiency of the repair system (Britten, 1986), generation time (Li et al., 1987), body size and metabolic rate (Martin

et al., 1993), and relaxed selective constraints (Adachi et al., 1993) are among them.

It is obvious that the generation time hypothesis does not fit to some mitochondrial DNA, because it cannot explain the faster evolutionary rates in lamprey and human lineages than in fishes and cow respectively. The body size of lamprey is much bigger than that of loach and carp and loach are very different in body size, which means that the body size hypothesis also cannot explain the heterogeneous rate of evolution in the mitochondrial system. Is the metabolic rate of fishes slower than those of frog and lamprey? Do poikilotherms have a stricter DNA repair system than homeotherms? These questions have not been answered. Probably as more data accumulate, additional exceptions will be uncovered. More than one hypothesis may be necessary to explain the evolutionary patterns.

Wilson et al. (1987) argued against the hypothesis of the faster rate of evolution in rodent lineage, since it was likely that the discrepancy of the rates between lineages was due to the uncertainty of fossil records. It also proves to be true in this study that the difference of evolutionary rate can be greatly overestimated, if fossil records are misinterpreted. The discrepancy in evolutionary rate between cold- and warm-blooded vertebrates can be from twice to five times depending on the use of the minimum or maximum diverged time from available fossil records.

In the previous studies, both sequence saturation and poor fossil records apparently contributed to the overestimation of the discrepancy in the evolutionary rate between the two major vertebrate groups. Although the JTT-F model is superior to

any other models available, it is surely far from the actual amino acid transition matrix. Thus, more accurate models for amino acid sequence substitutions in animal mitochondria, and additional comparisons of close relatives are needed to see how much faster the mtDNA of homeotherms evolve. To this end, sequences of hagfish and shark mitochondrial genomes may be useful.

Acknowledgement

I would like to thank Dr. Masami Hasegawa for providing the PROTML program and helpful discussion and comments on the results of this study. I also greatly appreciate Dr. Will Gilbert for his help in compiling PROTML.

LITERATURE CITED

Adachi, J. and M. Hasegawa. 1992. Molphy: Program for molecular phylogenetics, I. PROTML: Maximum likelihood inference of protein phylogeny. Institute of Statistical Mathematics, Tokyo.

Adachi, J., Y. Cao, and M. Hasegawa. 1993. Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid sequence level: Rapid evolution in warm-blooded vertebrates. J. Mol. Evol. 36:270-28.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Anderson, S., M.H.L. De Brujin, A.R. Coulson, I.C. Eperon, F. Sanger, and I.G. Young. 1982. Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. J. Mol. Biol. 156:683-717.

Anderson, S., A.T. Bankier, B.G. Barrell, M.H.L. de Bruijin, A.R. Coulson, J. Drouin, I.C. Eperon, D.P. Nierlich, B.A. Roe, F. Sanger, P.H. Schreier, A.J.H. Smith, R. Staden, and I.G. Young. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457-465.

Árnason, E., and D.M. Rand. 1992. Heteroplasmy of short tandem repeats in mitochondrial DNA of Atlantic cod, *Gadus morhua*. Genetics 132:211-220.

Árnason, Ú., A. Gullberg, and B. Widegren. 1991. The complete nucleotide sequence of the mitochondrial DNA molecule of the fin whale, *Balaenoptera physalus*. J. Mol. Evol. 33:556-568.

Avise, J.C. 1991. Ten unorthodox perspectives on evolution prompted by comparative population genetic findings on mitochondrial DNA. Ann. Rev. Genet. 25:45-69.

Avise, J.C., R.M. Ball and J. Arnold. 1988. Current versus historical population sizes in vertebrate species with high gene flow: A comparison based on mitochondrial DNA lineages and inbreeding theory for neutral mutations. Mol. Biol. Evol. 5:331-344.

Avise, J.C., G.S. Helfman, N.C. Saunders, and L.S. Hales. 1986. Mitochondrial DNA differentiation in North Atlantic eels: Population genetics consequences of an unusual life history pattern. Proc. Natl. Acad. Sci. USA. 83:4350-4354.

Barnes, W. 1977. Plasmid detection and sizing in single colony lysates. Science. 195:393.

Bartlett, S.E., and W.S. Davidson. 1991. Identification of *Thunnus* tuna species by the polymerase chain reaction and direct sequence analysis of their mitochondrial cytochrome b genes. Can. J. Fish. Aquat. Sci. 48:309-317.

Beckenbach, A.T. 1991. Rapid mtDNA sequence analysis of fish populations using the polymerase chain reaction (PCR). Can. J. Fish. Aquat. Sci. 48 (Suppl.1):95-98.

Bermingham, E., T. Lamb, and J.C. Avise. 1986. Size polymorphism and heteroplasmy in the mitochondrial DNA of lower vertebrates. J. of Heredity 77:249-252.

Bibb, M.J., R.A. Etten, C.T. Wright, M.W. Walberg and D.A. Clayton. 1981. Sequence and gene organization of mouse mitochondrial DNA. Cell 26:167-180.

Billington, N. and P.D. Hebert. 1991. Mitochondrial DNA diversity in fishes and its implications for introductions. Can. J. Fish. Aquat. Sci. 48(Suppl. 1):80-94.

Birky, Jr., C.W. 1991. Evolution and population genetics of organelle genes: Mechanisms and models. In Selander R.K., A.G. Clark, and T.S. Whittam (eds.), Evolution at the molecular level. Sinauer Associates. MA.

Birky, Jr., C.W., P. Fuerst, and T. Maruyama. 1989. Organelle gene diversity under migration, mutation, and drift: Equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. Genetics 121:613-627.

Birky, Jr., C.W., T. Maruyama, and P. Fuerst. 1983. An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. Genetics 103:513-527.

Britten, R.J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393-1398.

Brown, G.G., G. Gadaleta, G. Pepe. C. Saccone, and E. Sbisa. 1986. Structural conservation and variation in the D-loop-containing region of vertebrate mitochondrial DNA. Mol. Biol. 192:503-511.

Brown, G.G. and M.V. Simpson. 1982. Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. Proc. Natl. Acad. Sci. USA. 79:3246-3250.

Brown, J.R., A.T. Beckenbach, and M.J. Smith. 1992. Mitochondrial DNA length variation and heteroplasmy in populations of white sturgeon (*Acipenser transmontanus*). Genetics 132:221-228.

Brown, W.M. 1985. The mitochondrial genome of animals. In R.J. MacIntyre (ed.), Molecular evolutionary genetics. Plenum Press, New York.

Brown, W.M., M. George, Jr., and A.C. Wilson. 1979. Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA. 76(4):1967-1971.

Bucklin, A., B.W. Frost, and T.D. Kocher. 1992. DNA sequence variation of the mitochondrial 16S rRNA in *Calanus* (Copepoda; Calanoida): intra- and interspecific patterns. Molec. Mar. Biol. Biotech. 1(6):397-407.

Buroker, N.E., J.R. Brown, T.A. Gilbert, P.J. O'Hara, A.T. Beckenbach, W.K. Thomas and M.J. Smith. 1990. Length heteroplasmy of sturgeon mitochondrial DNA: an illegitimate elongation model. Genetics 124:157-63.

Cao, Y., J. Adachi, A. Janke, S. Pääbo, and M. Hasegawa. 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins. Submitted to J. Mol. Evol.

Cabot, E.L. and A.T. Beckenbach. 1989. Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. Comput. Appl. Biosci. 5:233-234.

Cann, R.L. and A.C. Wilson 1983. Length mutations in human mitochondrial DNA. Genetics 104:699-711.

Cantatore, P., M. Gadaleta, M. Robert, C. Saccone, and A.C. Wilson. 1987. Duplication and remoulding of transfer RNA genes during the evolutionary rearrangement of mitochondrial genomes. Nature 32:853-855.

Carr, S.M. and H.D. Marshall. 1991. A direct approach to the measurement of genetic variation in fish populations:applications of the polymerase chain reaction to studies of Atlantic cod, *Gadus morhua* L.. J. Fish. Bio. 39(A):101-107.

Carr, S. M. and H.D. Marshall. 1991. Detection of intraspecific DNA sequence variation in the mitochondrial cytochrome b gene of Atlantic cod (*Gadus morhua*) by the polymerase chain reaction. Can. J. Fish. Aquat. Sci. 48:48-52.

Carroll, R.L. 1988. Vertebrate paleontology and evolution. W.H. Freeman and Company, New York.

Chapman, R.W. and D.A. Powers. 1984. A method for the rapid isolation of mitochondrial DNA from fish. Technical report, Maryland sea grant program.

Chang, D.D. and D.A. Clayton. 1986. Identification of primary transcription start sites of mouse mitochondrial DNA: Accurate in vitro initiation of both heavy- and light-strand transcriptions. Mol. Cell. Biol., 6:1446-1453.

Chang Y.-S., F.-L. Huang, T.-B. Lo. 1994. The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. J. Mol. Evol. 38:138-155.

Chomyn, A. and G. Attardi. 1987. Mitochondrial gene products. Current Topics in Bioenergetics. Vol. 15:295-329.

Clary, D.O. and D.R. Wolstenholme. 1985. The mitochondrial DNA molecule of *Drosophila yakuba*: Nucleotide sequence, gene organization, and genetic code. J. Mol. Evol. 22:252-271.

Clayton, D.A. 1987. Nuclear gene products that function in mitochondrial DNA replication. Phil. Trans. R. Soc. Lond. B. 317:473-482.

Colbert, E.H. and M. Morales. 1992. Evolution of the vertebrates. A history of the backboned animals through time. Wiley-Liss, New York.

Crozier, R.H. and Y.C. Crozier. 1993. The mitochondrial genome of the honeybee *Apis mellifera*: Complete sequence and genome organization. Genetics 133:97-117.

Dagert, M. and S.D. Ehrlich. 1979. Prolonged incubation in calcium chloride improves the competence of *Escherichia coli*. Gene. 6:23.

Danzmann, R.G. M.M. Ferguson, S. Skúlason, S.S. Snorrason and D.L. G. Noakes. 1991. Mitochondrial DNA diversity among four sympatric morphs of Arctic charr, *Salvelinus alpinus* L., from Thingvallavatn, Iceland. J. Fish Biol. 39:649-659.

Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt. 1978. A model of evolutionary change in proteins. In M. O. Dayhoff (ed.): Atlas of protein sequence and structure, vol. 5, suppl. 3, Pp. 545-572, National Biomedical Research Foundation, Washington DC.

Doda, J.N., C.T. Wright, and D.A. Clayton. 1981. Elongation of displacement-loop strands in human and mouse mitochondrial DNA is arrested near specific template sequences. Proc. Natl. Acad. Sci. USA 78(10):6116-6120.

Desjardins, P., and R. Morais. 1990. Sequence and gene organization of the chicken mitochondrial genome. J. Mol. Biol. 212:599-634.

Dessauer, H.C., C.J. Cole, and M.S. Hafner. 1990. Collection and storage of tissues. In D.M. Hillis and C. Moritz (eds.), Molecular systematics. Sinauer Associates, Inc. MA, USA.

Dowling, T.E., C. Moritz, and J.D. Palmer. 1990. Nucleic acids II: Restriction site analysis. In D.M. Hillis and C. Moritz (eds.). Molecular systematics. Sinauer Associates, Inc. MA, USA.

Emanuel, J.R. 1991. Simple and efficient system for synthesis of non-radioactive nucleic acid hybridization probes using PCR. Nucleic Acids Research. 19(10):2790

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368-376.

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22:521-565.

Finnerty, J.R. and B.A. Block. 1992. Direct sequencing of mitochondrial DNA detects highly divergent haplotypes in blue marlin (*Makaira nigricans*). Mol. Mar. Biol. Biotech. 1:206-214.

Forey, P. and P. Janvier. 1993. Agnathans and the origin of jawed vertebrates. Nature 361:129-134.

Fukuda, M., S. Wakasugi, T. Tsuzuki, H. Nomiyama, and K. Shimada. 1985. Mitochondrial DNA-like sequences in the human nuclear genome. Characterization and implications in the evolution of mitochondrial DNA. J. Mol. Biol. 186:257-266.

Gadaleta, G., G. Pepe, G. de Candia, C. Quagliariello, E. Sbisa, and C. Saccone. 1989. The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: Cryptic signals revealed by comparative analysis between vertebrates. J. Mol. Evol. 28:497-516.

Garesse, R. 1988. *Drosophila melanogaster* mitochondrial DNA: Gene organization and evolutionary considerations. Genetics 118:649-663.

Gatesy, J., D. Yelon, R. Desalle, and E.S. Vrba. 1992. Phylogeny of the Bovidae (Artiodactyla, Mammalia), based on mitochondrial ribosomal DNA sequences. Mol. Biol. Evol. 9(3):433-446.

Gauldie, R.W. 1991. Taking stock of genetic concepts in fisheries management. Can. J. Fish. Aquat. Sci. 48:722-731.

Gorr, T. and T. Kleinschmidt. 1993. Evolutionary relationships of the coelacanth. Ameri. Scienti. 81:72-82.

Gorr, T., T. Kleinschmidt, and H. Fricke. 1991. Close tetrapod relationships of the coelacanth *Latimeria* indicated by haemoglobin sequences. Nature 351:394-397.

Graves, J.E., J.R. McDowell, A.M. Beardsley and D.R. Scoles. 1992. Stock structure of the bluefish *Pomatostomus saltatrix* along the mid-Atlantic coast. Fish. Bull. 90:703-710.

Gyllensten, U., D. Wharton, A. Josefsson, and A.C. Wilson. Paternal inheritance of mitochondrial DNA in mice. Nature 352:255-257.

Gyllensten, U., D. Wharton, and A.C. Wilson. 1985. Maternal inheritance of mitochondrial DNA during backcrossing of two species of mice. J. of Heredity 76:321-324.

Hannhan, D. 1985. Techniques for transformation of *E.coli*. In D.M.Glover (ed.),

DNA cloning: a practical approach. vol. 1. Pp.109-135. IRL Press, Oxford.

Henikoff, S. 1984. Unidirectional digestion with exonuclease III creates targeted break points for DNA sequencing. Gene. 28:351-359.

Hasegawa, M., A.D. Rienzo, T.D. Kocher, A.C. Wilson. 1993. Toward a more accurate time scale for the human mitochondrial DNA tree. J. Mol. Evol. 37:347-354.

Hasegawa, M. and M. Fujiwara. 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. Mole. Phyl. Evol. 2(1):1-5.

Hillis, D.M., M.T. Dixon, and L.K. Ammerman. 1991. The relationships of the coelacanth *Latimeria chalumnae*: Evidence from sequences of vertebrate 28S ribosomal RNA genes. Environ. Biol. Fishes 32:119-130.

Hoffmann, R.J., J.L. Boore, and W.M. Brown. 1992. A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. Genetics 131:397-412.

Hoelzel, A.R. 1993. Evolution by DNA turnover in the control region of vertebrate mitochondrial DNA. Current opinion in genetics and development 3:891-895.

Hoelzel, A.R., J.M. Hancock, and G.A. Dover. 1991. Evolution of the cetacean mitochondrial D-loop region. Mol. Biol. Evol. 8(3):475-495.

Hubbs, C.L. and I.C. Potter. 1971. Distribution, phylogeny and taxonomy. In Hardisty, M.W. and I.C. Potter (eds.). The biology of lampreys. Academic Press, New York.

Hutchinson III, C.A., J.E. Newbold, S.S. Potter, and M.H. Edgell. 1974. Maternal inheritance of mammalian mitochondrial DNA. Nature 251:536-538.

Irwin, D.M., T.D. Kocher, and A.C. Wilson. 1991. Evolution of the cytochrome b gene of mammal. J. Mol. Evol. 32:128-144.

Jacobs, H.T., D.J. Elliott, V.B. Math, and A. Farquharson. 1988. Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. J. Mol. Bio. 202:185-217.

Jamieson, B.G.M. 1991. In Fish evolution and systematics: Evidence from spermatozoa. Agnatha. Class Myxini. Order Myxiniformes. Class Cephalaspidomorphi. Order Petromyzontiformes. Cambridge University Press.

Janke, A., G. Feldmaier-Fuchs, W.K. Thomas, A.v. Haeseler, and S. Pääbo. 1994. The marsupial mitochondrial genome and the evolution of placental mammals. Genetics (in press).

Johansen, S., P.H. Guddal, and T. Johansen. 1990. Organization of the mitochondrial genome of Atlantic cod, *Gadus morhua*. Nucleic Acids Research. 18(3):411-419.

Johnson, M.J., D.C. Wallace, S.D. Ferris, M.C. Rattazzi, and L.L. Cavalli-Sforrza. 1983. Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. J. Mol. Evol. 19:255-271.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111-120.

Kocher, T.D., W.K. Thomas, A. Meyer, S.V. Edwards, S. Pääbo, F.X. Villablaca, and A.C. Wilson. 1989. Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. USA 86:6196-6200.

Kocher, T.D., and A.C. Wilson. 1991. Sequence evolution of mitochondrial DNA in human and chimpanzees. In S. Osawa and T. Honjo (eds.). Evolution of life: fossils, molecules and culture. Pp. 391-413. Springer-Verlag, Tokyo.

Kocher, T.D. 1992. PCR, direct sequencing, and the comparative approach. PCR: methods and applications 1(4):217-221.

Kondo, R., Y. Satta, E.T. Matsuura, H. Ishiwa, N. Takahata, and S.I. Chigusa. 1990. Incomplete maternal transmission of mitochondrial DNA in *Drosophila*. Genetics 126:657-663.

Kumar, S., K. Tamura, and M. Nei. 1993. Mega: Molecular evolutionary genetics analysis, Ver. 1.0. The Pennsylvania State University, University Park, PA 16802.

Kumazawa, Y., M. Nishida. 1993. Sequence evolution of mitochondrial tRNA

genes and deep-branch animal phylogeny. J. Mol. Evol. 37:380-398.

Lake, J.A. 1987. Determining evolutionary distances from highly diverged nucleic acid sequences: Operator metrics. J. Mol. Evol. 26:59-73.

Lansman, R.A., R.O. Shade, J.F. Shapira, and J.C. Avise. 1981. The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations: III. Techniques and potential applications. J. Mol. Evol. 17:214-226.

Lauder, G.V. and K.F. Liem, 1983. The evolution and interrelationships of the actinopterygian fishes. Bull. Mus. Comp. Zool. 150(3):95-197.

Lee, W.-J., J. Conroy, W.H. Howell, and T.D. Kocher. 1994. Structure and evolution of teleost mitochondrial control regions. Submitted to J. Mol. Evol.

Levinson, G. and G.A. Gutman. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4:203-221.

Lin, J. and M. Nei. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Bio. Evol. 7(1):82-102.

Martin, A.P., G.J.P. Naylor, and S.R. Palumbi. 1992. Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. Nature 357:153-155.

Martin, A.P. and S.R. Palumbi. 1993. Body size, metabolic rate, generation time, and molecular clock. Proc. Natl. Acad. Sci. USA 90:4087-4091.

Meyer, A. and S.I. Dolven. 1992. Molecules, fossils, and the origin of tetrapods. J. Mol. Evol. 35:102-113.

Michaels, G.S., W.W. Hauswirth, and P.J. Laipis. 1982. Mitochondrial DNA copy number in bovine oocytes and somatic cells. Develop. Biol. 94:246-251.

Mignotte, B., D. Dunon-Bluteau, C. Reiss, and J-C. Mounolou. 1987. Sequence deduced physical properties in the D-loop region common to five vertebrate mitochondrial DNAs. J. Theor. Biol. 124:57-69.

Mignotte, F., M. Gueride, A.-M. Champagne, and J.-C. Mounolou, 1990. Direct repeats in the non-coding region of rabbit mitochondrial DNA. Eur. J. Biochem. 194:561-571.

Moritz, C., T.E. Dowling, and W.M. Brown. 1987. Evolution of animal mitochondrial DNA: Relevance for population biology and systematics. Ann. Rev. Ecol. Sys. 18:269-92.

Moyle, P.B. and J.J. Cech, Jr. 1988. Fishes. An introduction to ichthyology. Prentice-Hall, NJ.

Mulligan, T.J. and R.W. Chapman. 1989. Mitochondrial DNA analysis of Chesapeake Bay white perch, *Morone americana*. Copeia 1989(3):679-688.

Nolan, K. and J. Grossfield, and I. Wirgin. 1991. Discrimination among Atlantic coast populations of American shad (*Alosa sapidissima*) using mitochondrial DNA. Can. J. Fish. Aquat. Sci. 48:1724-1734.

Normark, B.B., A.M. McCune, and R.G. Harrison. 1991. Phylogenetic relationships of neopterygian fishes, inferred from mitochondrial DNA sequences. Mol. Biol. Evol. 8(6):819-834.

Okimoto, R., J.L. Macfarlane, D.O. Clay, and D.R. Wolstenholme. 1992. The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. Genetics 130:471-498.

Ojala, D., J. Montoya, and G. Attardi. 1981. tRNA punctuation model of RNA processing in human mitochondria. Nature 290:470-290.

Olivo, P.D., M.J.V. de Walle, P.J. Laipis, and W.W. Hauswirth. 1983. Nucleotide sequence evidence for rapid genotypic shift in the bovine mitochondrial DNA D-loop. Nature 306:400-402.

Ovenden, J.R. 1990. Mitochondrial DNA and marine stock assessment: A review. Aust. J. Mar. Freshwater Res. 41:835-853.

Peer, Y.V.d., J.-M. Neefs, and R.d. Watchter. 1990. Small ribosomal subunit RNA sequences, evolutionary relationships among different life forms, and mitochondrial origin. J. Mol. Evol. 30:463-476.

Perna, T.N. and T.D. Kocher. 1994. Patterns of nucleotide composition at four-fold degenerate sites of animal mitochondrial genomes. Submitted to J. Mol. Evol.

Powell, J.R. and M.C. Zúñiga. 1983. A simplified procedure for studying mtDNA polymorphisms. Biochem. Genet. 21:1051-1055.

Reeves, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. 35:17-31.

Roe, B.A., D.-P. Ma, R.K. Wilson, and J. F.-H. Wong. 1985. The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J. Biol.Chem. 260(17):9759-9774.

Saccone, C., G. Pesole, and E. Sbisa. 1991. The main regulatory region of mammalian mitochondrial DNA: Structure-function model and evolutionary pattern. J. Mol. Evol. 33:83-91.

Saccone, C., M. Attimonelli, and E. Sbisa. 1987. Structural elements highly preserved during the evolution of the D-loop-containing region in vertebrate mitochondrial DNA. J. Mol. Evol. 26:205-211.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4(4):406-425.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. Molecular cloning: a laboratory manual (2nd ed.). Cold Spring Harbor Laboratory Press, NY.

Shedlock, A.M., J.D. Parker, D.A. Crispin, T.W. Pietsch, and G.C. Burmer. 1992. Evolution of the salmonid mitochondrial control region. Mol. Phyl. Evol. 1:179-192

Shields, G.F. and T.D. Kocher. 1991. Phylogenetic relationships of North American ursids based on analysis of mitochondrial DNA. Evolution 45:218-221.

Smith, M.F., W.K. Thomas, and J.L. Patton. 1992. Mitochondrial DNA-like sequence in the nuclear genomic of an akodontine rodent. Mol. Bio. Evol. 9(2):204-215.

Smith, M.J., A. Arndt, S. Gorski, and E. Fajber. 1993. The phylogeny of echinoderm classes based on mitochondrial gene arrangements. J. Mol. Evol. 36:545-554.

Smith, M.J., D.K. Banfield, K. Doteval, S. Gorski, and D.J. Kowbel. 1989. Gene arrangement in sea star mitochondrial DNA demonstrates a major inversion event during echinoderm evolution. Gene 76:181-185.

Stock, D.W. and G.S. Whitt. 1992. Evidence from 18S ribosomal RNA sequences

that lampreys and hagfishes form a natural group. Science 257:787-789.

Snyder, M., A.R. Fraser, J. LaRoche, K.E. Gartner-Kepkay, and E. Zouros. 1987. Atypical mitochondrial DNA from the sea-scallop *Placopecten magellanicus*. Proc. Natl. Acad. Sci. USA 84:7595-7599.

Stewart, D.T. and A.J. Baker. 1994. Patterns of sequence variation in the mitochondrial D-loop region of shrew. Mol. Bio. Evol. 11(1):9-21.

Southern, S.O., P.J. Southern, and A.E. Dizon. 1988. Molecular characterization of a cloned dolphin mitochondrial genome. J. Mol. Evol. 28:32-42.

Takahata, N. and S.R. Palumbi. 1985. Extranuclear differentiation and gene flow in the finite island model. Genetics 109:441-457.

Tamura, K. 1992. The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol. Biol. Evol. 9(5):814-825.

Tzeng, C.-S., C.-F. Hui, S.-C. Shen, and P.C. Huang. 1992. The complete nucleotide sequence of the *Crossostoma lacustre* mitochondriual genome: conservation and variation among vertebrates. NAR 20(18):4853-4858.

Vawter, L. and W.M. Brown. 1993. Rates and patterns of base change in the small subunit ribosomal RNA gene. Genetics 134:597-608.

Warrior, R., and J. Gall. 1985. The mitochondrial DNA of *Hydra attenuata* and *Hydra littoralis* consists of two linear molecules. Arch. Sci. Geneva 38:439-445.

Welter, C., S. Dooley, and N. Blin. 1989. A rapid protocol for the purification of mitochondrial DNA suitable for studying restriction fragment length polymorphisms. Gene. 83:169-172.

Wheeler, W.C. and R.L. Honeycutt. 1988. Paired sequence difference in ribosomal RNAs: Evolutionary and phylogenetic implications. Mol. Biol. Evol. 5(1):90-96.

Wiesner, R.J., H. Swift, and R. Zak. 1991. Purification of mitochondrial DNA from total cellular DNA of small tissue samples. Gene. 98:277-281.

Wilkinson, G.S. and A.M. Chapman. 1991. Length and sequence variation in evening bat D-loop mtDNA. Genetics 128:607-617.

Wilson, A.C., H. Ochman, and E.M. Prager. 1987. Molecular time scale for evolution. Trends in Genetics 3(9):241-247.

Wilson, A.C., R.L. Cann, S.M. Carr, M. George, U.B. Gyllensten, K.M. Helm-Bychowski, R.G. Higuchi, S.R. Palumbi, E.M. Prager, R.D. Sage, and M. Stoneking. 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. Biol. J. Linnean Soc. 26:375-400.

Wilson Jr, R.R., M.D. Tringail. 1990. Improved methods for isolation of fish mtDNA by ultracentrifugation and visualization of restriction fragments using fluorochrome dye: results from Gulf of Mexico clupeids. Fish. Bull., U.S. 88:611-615.

Wong, J.F., E.P. Ma, R.K. Wilson, and B.A. Roe. 1983. DNA sequence of the *Xenopus laevis* mitochondrial heavy and light strand replication origins and flanking tRNA genes. Nucl. Acids Res. 11(14):4977-4995.

Xiong, B. and T.D. Kocher. 1990. Comparison of mitochondrial DNA sequences of seven morphospecies of black flies (Diptera:Simuliidae). Genome 34:306-311.

Yokobori, S-i, T. Ueda, and K. Watanabe. 1993. Codons AGA and AGG are read as glycine in ascidian mitochondria. J. Mol. Evol. 36:1-8.

Zullo, S., L.C. Sieu, J.L. Slightom, H.I. Hadler, and J.M. Eisenstadt. 1991. Mitochondrial D-loop sequences are integrated in the rat nuclear genome. J. Mol. Evol. 221:1223-1235.